Phonetic Detail and Grammaticality Judgements

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of
Master of Arts
in Linguistics in the
University of Canterbury
by
Abby Walker

Contents

List of Figures

List of Tables

Acknowledgements

Firstly, thanks to the participants of all (4) experiments, and to the lovely folk who provided stimuli. Thanks also to the New Zealand Federation of Graduate Women (NZFGW) for a prize that enabled me to pay the participants, and to my wonderful Department of Linguistics for some funding magic that allowed me to present part of this thesis at ICPhS 16, in Saarbrücken. This thesis was funded by a University of Canterbury Masters Scholarship.

To Jen Hay, my supervisor and mentor, thanks so much, for the advice, direction, support, opportunities and just general linguistic inspiration. I owe you many beers.

To Heidi Quinn, my co-supervisor, syntax advisor and personal cheerleader, thanks also, and for marking this too!

To Paul Warren, for marking this thesis and thoughtful comments.

To my ling big sister, Katie Drager, who inspires me and motivates me and plain helps me, thank you.

To Rui Chaves and Anita Szakay for the LaTeX help. Cheers.

To the SocioKaffee Klatsch gang at Canterbury, for all the advice and help and *love*.

To Elizabeth Lochhead, for tolerance and support.

And lastly, to Mum and Dad, who I never thank as much as I ought, but who are behind every good thing I do.

Abstract

This thesis investigates predictions of an exemplar account of syntax, by testing whether manipulating socially salient phonetic detail can alter the grammaticality judgements given to morpho-syntactic constructions in New Zealand English (NZE).

Three experiments are were conducted as part of this thesis. The first tested the social saliency of different phonetic variables in NZE, and found phrase final /t/, which can be realised with or without a release, to be strongest. In the second experiment, phrase final /t/ was tested further, and manipulating the release significantly altered both the age and class ratings given to speakers. The way in which it did this reflected the patterns documented in production.

In the third experiment, participants were asked to rate the grammaticality of the same sentences. When the results of the previous experiment were included in the statistical model, an effect of the variant came out as significant. The more participants had rated a speaker as older with the released variant in the previous experiment, the less they rated the sentence as grammatical with the released variant. That is, only the most socially salient realisations were able to alter perceived grammaticality.

Overall, the results of this thesis suggest that speaker information and phonetic detail can affect grammaticality judgements. This supports an exemplar model of syntax. Regardless of the theoretical implications of the findings however, the methodological ones are clear. If speakers and realisations of certain phonetic variables can alter grammaticality judgements, then they must be controlled for in the presentation of stimuli to participants.

Chapter 1

Introduction

The production and processing of morpho-syntactic constructions involve speakers and speech. While traditional models of syntax would relegate these factors outside of the language faculty (cf. Chomsky 1995), the emergence of usage-based models opens up the possibility that such factors contribute to the grammar itself. In this thesis I explore the predictions of an exemplar model of syntax. Exemplar Theory (ET) proposes that encountered instances of speech are stored as richly detailed and annotated memories (Johnson 1997, Pierrehumbert 2006). Such episodic models have been successfully used in phonology to account for otherwise problematic experimental findings, such as the fact that listeners categorise the same incoming signal differently depending on speaker-related information (Strand 1999, Strand & Johnson 1996, Niedzielski 1999, Drager 2005, Hay, Warren & Drager 2006).

Applying such a model to syntax would predict that morpho-syntactic constructions are stored with phonetic and speaker information. If a construction is used more by a particular group in society, then the memory of that construction will have a more robust association with those speakers. The acceptability or grammaticality of that construction may then become dependent on whether the speaker belongs to this group or not. And as certain social groups also use certain phonetic variants more than other groups, it is possible that the acceptability or grammaticality of that construction may even be dependent on the realisation of phonetic variables[1].

[1]It is the association of speaker information with a *type* of construction that is of primary concern here, not any particular instances or *tokens* of constructions, though this would also be worth pursuing.

The explicit aim of this thesis is to test whether socially meaningful phonetic detail can alter the grammaticality judgements given to socially variable constructions. Three experiments were conducted. The first two tested the social saliency of different phonetic variables in New Zealand English (NZE), and the third tested whether the manipulation of one of these, phrase final /t/, could alter the perceived grammaticality of sentences.

The structure of this thesis is as follows:

- In Chapter 2 I outline the relevant background information for the thesis, which incorporates a range of linguistic disciplines. I begin with a review of ET, focusing on some recent speech perception experiments, then turn to usage-based models of syntax, and discuss applying ET at this level. Such a model would predict that encountered instances of constructions are stored with speaker information and even rich phonetic detail. The aim of the thesis is to find evidence that this is the case, and as my methodology involves socially salient phonetic detail, and grammaticality judgements (GJs), these are also discussed in some detail.

- In Chapters 3 and 4, I describe two experiments that tested the social saliency of certain phonetic variables in NZE. In the Pilot Study (Chapter 3), several candidate variables were tested, and the results suggested that the manipulation of phrase final /t/ was best able to alter the perceived age and class of speakers. In Chapter 4, I discuss how the social saliency of phrase final /t/ was further tested, this time embedded in the same sentences that were used in the GJ Experiment. The methodology and results of the GJ Experiment are presented in Chapter 5.

- In Chapter 6, I discuss the implications of my results. In terms of syntactic theory, I argue that my results pose problems for traditional models of syntax, but are compatible with an exemplar model of syntax. Methodologically, my results show that phonetic detail and speaker information can affect GJs, and therefore should be controlled for in the presentation of judgement stimuli. I also make an argument for including speaker information in the calculation of frequencies and probabilities. The chapter ends with an outline of a number of potential

follow up studies. In Chapter 7 I summarise my findings and the thesis contribution.

Chapter 2

Background

This thesis describes a study designed to test whether phonetic detail can alter grammaticality judgements. The hypothesis behind the experiment is that encountered morpho-syntactic constructions are stored with speaker information. Phonetic detail is used as a way to represent, in a controlled manner, such speaker information, and in an exemplar model would also be stored with a construction. GJs are used as a means of accessing the internal grammar.

The hypothesis is not pulled from thin air, but builds, as one of the next logical steps, on recent literature concerning episodic models of speech production and perception, and probabilistic models of syntax. These will be discussed in sections 2.1 and 2.2 respectively. In section 2.3, I review some previous studies suggesting that phonetic detail contributes to perceived speaker attributes. In section 2.4, I discuss GJs in terms of why they are used, how they are gathered, and what they are thought to represent. In section 2.5 I outline the aims of this thesis.

2.1 Episodic models of language

Exemplar Theory (ET) was first developed in psychology (Hintzman 1986, Nosofsky 1986), and brought to linguistics through the works of Johnson (1997) and Goldinger (1996) as a means to better explain how language users correctly categorise and process the wide and systematic phonetic variation they are exposed to. In exemplar models, encountered instances of speech

are stored as complete memories, rich with phonetic detail and any other information that the listener finds salient at the time, such as attributes of the speaker. This is in high contrast with standard, abstracted models of language. As Keith Johnson puts it:

> "*...the key idea of the exemplar-based approach is that people remember, at the core of the cognitive representation of language, linguistic episodes, not linguistic descriptions. We operate from mental images - detailed memories of specific linguistic experiences - rather than from impoverished descriptions of such experience*" (Johnson 2001, 492).

While simple exemplar models, such as Johnson's XMOD (1997) posit no explicit categories, Pierrehumbert (2006) makes a strong argument for a hybrid model of concrete phonetics and abstracted phonology, so that there are multiple levels of representation. Each category "is represented in memory by a large cloud of remembered tokens of that category" (Pierrehumbert 2001, 3). That is, the abstracted category is defined by the real and concrete memories, or exemplars, of tokens that were coded as belonging to that category during speech processing. What constitutes an exemplar can range from a sound approximating a phoneme to whole words, frequent collocations, phrases and even constructions. They are organized in the mind such that similar exemplars are closer together and dissimilar ones are further apart (Pierrehumbert 2001). This allows for overlap in production and ambiguity in perception. It can also explain why words with high neighborhood densities take longer to process than words with low neighborhood densities, due to the activation of the close and competing neighbouring words (Vitevitch & Luce 1998, Vitevitch & Luce 1999).

Phoneme (and word) categorization, according to the model, involves "comparing the to-be-categorized item with each of the remembered instances of each category, and categorization is based on sums of similarity over each category" (Johnson 1997, 146). A match is made when a potential candidate crosses a match threshhold. Every exemplar does not compete equally for the match, as "(m)ore proto-typical or central exemplars will be easy to access, because of their high resting activation level" (Mendoza-Denton et al. 2003, 134). The context can also activate certain exemplars, so it is not only the similarity of the phonetic signal that is important, but of the

entire memory, including any contextual information: "the model makes it possible to categorize new items by reference to *appropriate* prior examples - a subset of exemplars that resemble the to-be-recognized item on speaker specific dimensions" (Johnson 1997, 148, emphasis added). It is this feature of the model that renders speaker normalization unnecessary, because in activating exemplars for the relevant speaker, the model automatically alters categorization boundaries for the incoming signal. These boundaries will reflect previous experience of that particular speaker group, and even of that particular individual.

Production, Pierrehumbert (2001) suggests, consists of selecting a target category and then selecting at random a target location inside the categorys 'cloud', so that a number of exemplars contribute to production, forcing an entrenchment-style reversion back to the mean. In the production-perception loop, this process is tempered by other factors, leading to variation and change. Noise, in the form of articulatory deviations, can cause the area covered by a category label to spread. Systematic production biases can cause the category to shift toward a certain realization.

An exemplar theory of language crucially relies on a powerful memory capacity, and Johnson cites some studies in psychology suggesting that this is indeed something we possess (Johnson 1997, 147). He also suggests, using a model developed by Kruschke (1992), that rather than literally storing each exemplar, we map our experiences onto a connectionist map of the auditory space, so that instead of having 50 separate but identical exemplars, we have one point on the auditory map that has an association weight of 50 to a certain category label (it could, of course, also have association weights to other labels). Pierrehumbert (2001) also suggests that the memories are granularised, such that differences between exemplars that are minute are not recorded. An incoming token that essentially matches, at this granularised level, a pre-existing exemplar can simply reinforce it rather than create a new memory. She also points out that memories fade, so that more recent exemplars are stronger than older ones.

Intrinsic in the system is both inter and intra speaker variation, a phenomenon that has long been ignored by phonologists (and syntacticians) as being a product of performance, and as not reflecting the underlying competence of speakers. Gleitman and Gleitman expressed the sentiments of

many when they said if we could strip away various contaminating factors in behavior, we might see the grammar bare (Gleitman & Gleitman 1970, 10). In ET, however, variation is part of competence and performance - indeed, one could argue that the two are one and the same. The model stores the variation that language users encounter and is dependent on that variation for speech processing. Embracing synchronic variation in a dynamic system also smoothly accounts for diachronic change, without the need for a system reconfiguration.

ET explains the well attested frequency effects seen in all aspects of language use. Frequency comes for free in an exemplar model: more frequent words or categories have more exemplars and more highly activated exemplars than infrequent words or categories. This explains frequency effects seen in production, where more frequent words are produced further along in leniting sound changes than less frequent words (cf. Hooper 1976, Phillips 1998). Systematic production biases, or 'phonetic rules' (Bybee 1994) work on each individual category/word each time it is used, so that shifts will happen fastest in the categories/words that are used most (Pierrehumbert 2001). At the same time, ET can also account for the initially conflicting phenomenon that sees frequent words being less prone to change by analogy, and other similar 'nonphonetic rules', than infrequent words (Hooper 1976, Phillips 1998, Phillips 2001). Frequent words have entrenched and specific forms. When a general, synthesized rule sweeps the lexicon, the existing exemplars will be too robust for the category to be affected. Infrequent words, with their impoverished set of exemplars, will quickly regularise (Zuraw 2003, 159).

In perception, frequent words are accessed (Oldfield & Wingfield 1965) and processed (cf. Monsell, Doyle & Haggard 1989) more quickly than infrequent words. This is easily accounted for in ET, as frequent words have higher resting activation levels than infrequent words, making them more easily accessible.

2.1.1 Exemplar Theory and speaker information

Another major strength of ET is that, by dictating that typically extralinguistic information is encoded in the exemplars, it can account for a number of recent and interesting experimental findings regarding speech perception

and speaker identity. In exemplar models "social information that is interpretable by a listener is automatically stored with the exemplar, made more robust with repetition, and crucially linked to the actual instances of use of a particular variant" (Mendoza-Denton et al. 2003, 136). This means that listeners rely on speaker information in categorizing incoming sounds[1],and thus manipulating perceived speaker attributes can change listeners perception of the physical signal (Strand 1999, 87).

Strand and Johnson (1996) did just this when they presented participants with audiovisual stimuli[2]. Using tokens from /s/ and /ʃ/ continuums constructed using non-prototypical male and female voices, participants were asked to say whether they heard the word *sod* or *shod*. This auditory channel was matched with a video of either prototypically male or female faces saying *sod*. Half of the tokens were gender matched audiovisually (male voice, male face), and half were mismatched (male voice, female face). Their results showed that the same recording could be perceived differently depending on the sex of the face it was presented with, and the boundary shift reflected differences in the production of /s/ and /ʃ/ by men and women. This suggested that "representations of speech categories as well as stereotypes about gender, nationality, race, and so forth must necessarily be interconnected" (Strand 1999, 98). In a follow up study, Munson, McDonald, DeBoe & White (2006) showed that even the perceived sexual orientation of female speakers could affect the perceived phoneme, such that a token from a heterosexual female was more likely to be perceived as /ʃ/ than a token from a lesbian/bisexual sounding female.

Niedzielski (1999) achieved a related effect by manipulating the perceived dialect of the speaker. Participants in Detroit were asked to match a target vowel in a sentence to one of a continuum of synthesized MOUTH vowels. In one condition, participants were told that the speaker was Canadian; in the other condition, they were told that the speaker was from Detroit. The condition participants were in was a significant factor in which vowel they believed they heard: in the Canadian condition they were more likely to correctly identify the Canadian-raising in the stimuli than were those in the

[1]And, as I shown in Chapter 4, use phonetic detail in making judgements about speaker attributes

[2]See also Johnson, Strand & DImperio (1999) for a similar study looking at perceived vowels

Detroit condition. She concluded that "(l)isteners do use social information to calibrate the phonological space of speakers" (Niedzielski 1999, 84). A series of experiments (Hay, Nolan & Drager (2006), Hay & Drager (forthcoming), Hay, Walker & Drager (under review)) from the Origins of New Zealand English Project (ONZE) suggest that participants do not need to believe that a speaker has a different dialect for that dialect to alter the phonological space. Rather, they found that the mere exposure to the concept of a region caused similar effects in their subjects.

Hay, Warren & Drager (2006) (see also Drager 2005) showed that the perceived age of the speaker can affect whether participants can hear the distinction between two vowels that are merging. In NZE, the diphthongs NEAR and SQUARE are undergoing a merger, such that younger speakers have almost entirely merged on NEAR, while older speakers still maintain the distinction. Participants in a study were presented with recordings of NEAR and SQUARE words and photos of older or younger people. Participants who were not fully merged themselves were better able to correctly distinguish the vowels when they were accompanied by the photos of older speakers than when they were accompanied by photos of the younger speakers, showing "relatively sophisticated sensitivity to social factors" (Hay, Warren & Drager 2006, 479).

ET can not only account for these findings but actually predicts them. If encountered instances of speech are stored with speaker information, and different speakers realize certain variables differently, then it falls out that "(l)isteners normalize speech through reference to experience-based expectations regarding speaker-to-speaker variation" (Strand 1999, 89). How language is processed is critically reliant on who people are listening to, or at least, who they think they are listening to.

2.2 Usage-based models of syntax

In much the same tradition as phonologists, syntacticians have long considered language, at its core representation, to be strictly categorical and hard-wired. This has lead to a theoretical chasm between competence, being the underlying mechanisms controlling the grammar, and performance, being the concrete reality of language use. Generally, syntax as a field is concerned with "how to learn about the former based on the latter" (Schütze 1996,

21), and variation and gradation are considered to be results of irrelevant extra-grammatical processes. Joos (1966, 351) expressed this most strongly:

> "*...all phenomena, whether popularly regarded as linguistic... or not, which we find we cannot describe precisely with a finite number of absolute categories, we classify as non-linguistic elements of the real world and expel them from linguistic science. Let sociolinguists and others do what they like with their own terminology... that continuity which we refuse to tolerate in our own science*".

There is a growing movement, however, to embrace at least some of the variation seen in usage as being part of the linguistic system. Lakoff famously argued that "Fuzzy grammar has a mental reality" (Lakoff 1973, 286), and Hawkins that there is "a profound correspondence between performance and grammars" (Hawkins 2004, xi). A number of edited volumes have now been published with the primary aim of accounting for variation and gradation in syntax (Aarts, Denison, Keizer & Popova 2004, Fanselow, Fery, Schlesewsky & Vogel 2006). By not accounting for these phenomena, many feel that "there are facts about linguistic theory and about the grammars of a particular language whose existence will be obscured" (Elliot, Legum & Annear Thompson 1969, 52).

Such lines of reasoning have lead to an increasing interest in usage-based and probabilistic models of syntax, including the exploration of incorporating principles of ET into syntactic models (cf. Hay & Bresnan 2006, Bod 2006). The question is, if one assumes that encountered instances of language are stored, can this abstract beyond the concrete phonetic signal to the underlying and generative construction? That is, words and phonemes aside, can we see effects of usage and probabilities on the hidden structures of language?

Gahl and Garnsey (2004) set out to test this hypothesis explicitly by looking at verb biases in English. Certain verbs, like *confirm*, can take direct objects (DO), as in the sentence *John confirmed the date of his visit*, as well sentential complements (SC), as in the sentence *John confirmed that he would come on the 29th May*. However, corpus data show that this verb occurs most frequently with direct objects, and is thus said to be biased to taking

direct objects. Importantly, this bias does not concern the probability of particular words occurring next to each other, but rather, the probability of a verb being followed by a particular construction.

To test the hypothesis that the probability of a biased verb taking a construction is in the grammar, Gahl and Garnsey decided to look at the acoustic features of both verbs with DO and SC biases in both DO and SC constructions. As already mentioned in Section 2.1, more frequent words are produced more reduced than frequent words (for example, Hooper 1976), and so Gahl and Garnsey's hypothesis was that when the verbs occur in the more probable and bias-confirming construction, they are more reduced (shorter in length and with higher rates of t/d-deletion) than when they occur in the less probable and bias-violating construction. The sentences were heavily normed so that the combinations of words were equally frequent and the sentences of similar lengths. Their results confirmed their hypothesis, suggesting that "knowledge of syntactic probabilities is part and parcel of syntactic knowledge" (Gahl & Garnsey 2004, 766).

Hay and Bresnan (2006) looked at the phonetic realisation of the noun *hand* and the verb *give* in various contexts. They found that *hand* was more raised (further along in a sound change) when literally referring to the body part than when being used in figurative constructions such as *give a hand, lend a hand*, etc. *Give*, in contrast, was more centralised (further along in a sound change) when the object was abstract (*give me an idea*) rather than when a physical transfer was being described (*give me that book*). *Hand* occurs most frequently with its literal sense, while *give* occurs most frequently with its abstract meaning. Hay and Bresnan argue that the difference in behaviour of the two words might be because nouns have a more independent representation than verbs, which, being more restricted in the objects they can take, occur more frequently with certain object types and thus had strong and robust associations with those objects. Their results provide support "both to the idea that phrases may be stored, and to the idea that this storage may be phonetically detailed" (Hay & Bresnan 2006, 337).

Bod (2000) showed that participants, asked to decide whether a three word string was English or not, responded faster to frequent strings such as *I like it* than to infrequent strongs such as *I keep it*. His experiment controlled for semantic plausibility, lexical frequency and syntactic complexity, and the

frequent strings were non-idiosyncratic. He concludes from his results that "frequent sentences must somehow be stored in memory" Bod (2000, 1).

In Bresnan, Cueni, Nikitina & Baayen (2005), the probability of a particular realization of the dative alternation was shown to be dependent on a number of factors, and the probability of the prepostional complement occurring increased more as more of the factors aligned to favour it. Bresnan (2005) went on to show that based on the number of aligning factors in a sentence, participants could not only fairly accurately predict which form of the dative alternation had been used in the real speech of another, but could also "make accurate probabilistic predictions of the syntactic choices of others" (Bresnan 2005, 17). These results suggested that language users were extremely sensitive to syntactic probabilities.

Bybee, after considering her own evidence, summarised that "grammar (is) the cognitive organization of ones experience with language" (Bybee 2006, 711). She calls for a view of language "based on constructions and as having an exemplar representation in which specific instances of use affect representation" (Bybee 2006, 715). An ET account of morpho-syntax would predict and explain the frequency and probability effects described above, because frequency and probabilities are an intrinsic part of storage[3].

Such an account would also allow "specific information about instances of use to be retained in representation" (Bybee 2006, 718), including salient speaker information. This would predict, as we saw in 2.1 in the examples from phonology, that manipulating speaker information could also predict how certain constructions were perceived.

2.2.1 Are constructions stored with speaker information?

As with phonological variables, some morpho-syntactic constructions vary in a socially-meaningful way, such that one group of speakers may use a particular form more than another group (for NZE, see Quinn (2004)). If language users are storing constructions with relevant speaker information, we would

[3]It would also account for the data in support of continuous grammatical categories (see Manning 2003, Aarts 2007)

expect them to be sensitive to social information in the perception of these constructions.

In Walker (2005*b*), I discuss an experiment designed to see if this was the case. I recorded two male speakers reading a number of sentences that contained a range of morpho-syntactic constructions that showed social variation in their production in NZE, such that younger speakers were more likely to use them than older speakers. The two males differed primarily in their age, and one was around thirty years older than the other. Through the course of the experiment, participants would hear each sentence twice, once read by the younger speaker, and once read by the older speaker. They were asked to rate the grammaticality of each sentence on a six point scale (see 2.4 this chapter for a discussion of GJs).

For non-standard preterite constructions (see 4.1.1 for a discussion), there was an effect of the speaker on the ratings of the least grammatical of the constructions[4], such that they received higher grammaticality ratings when spoken by the younger speaker. As the participants would have encountered these constructions more frequently coming from younger speakers, they would have a more robust representation of the construction tagged with young speakers than with older speakers. If the grammaticality judgements were in part a reflection of frequency, then these results would be congruent with just such an explanation.

While these results suggest that there is at least an association of certain speakers with certain constructions in peoples language faculties, it is not clear that the differences were entirely due to the ages of the speakers. One of the speakers might have sounded friendlier, or read the sentences more naturally, or with a prosody that worked with the constructions better. There were simply too many potential variables in the recordings.

Thus, in this thesis, I set out to run a similar experiment to the one I did in 2005, again testing whether the social attributes of a speaker can alter the grammaticality judgements given to non-standard constructions. This time, however, the compared recordings would differ only in one or two socially meaningful phonetic variables, so that if there were a difference, it would be

[4]Based on written ratings of the sentences

much easier to claim that it is due to the social associations of the variable alone.

2.3 The social saliency of phonetic variables

It is well documented that the phonetic realization of sociolinguistic variables can systematically differ in production according to the social attributes of a speaker, such as their age, class or ethnicity (cf. Labov 1972, Trudgill 1974). What is less understood is the degree to which listeners routinely exploit this systematicity in order to make social judgments about speakers.

Studies show that listeners are fairly accurate in making judgments about a speakers identity, whether it be identifying a particular person or simply identifying that the speaker belongs to a certain group in society. General research into what listeners use to do this have mainly focussed on voice quality features, and Van Lancker, Kreiman and Wickens (1985), looking at the recognition of famous voices in normal and reversed speech, conclude that recognition comes from "pitch, pitch range, rate, vocal quality, and vowel quality, but without benefit of acoustic detail reflecting specific articulatory and phonetic patterns, and orderly temporal structure" (Van Lancker et al. 1985, 30). However, there are a handful of studies that suggest that specific articulatory and phonetic patterns can in fact influence speaker recognition.

Participants in a study by Remez, Fellowes and Rubin (1997) could recognise colleagues from sinewave speech, where the original, normal speech recordings had been modified by sinusoidal (pure tone) replication. Such a transformation removes features of voice quality like pitch and intonation, but phonetic detail remains, and most listeners are still able to correctly understand the linguistic content. Their positive results heavily suggested that speakerspecific phonetic realisations, or idiolects, were remembered and used by listeners in speaker identification. Sheffert, Pisoni, Fellows & Remez (2002), in a series of experiments following up the study, showed that participants trained to recognise different speakers from sinewave speech could apply this knowledge to novel sentences not only from sinewave signals, but equally well, if not better, from natural speech. The combined results of the studies led researchers to conclude that "conceivably, identifying words and talkers could be based on a general capacity to discriminate the subtleties of phonetic expression"

(Sheffert et al. 2002, 1467). However, what specific phonetic cues listeners were using in the identification task was not examined.

Work by Purnell, Idsardi & Baugh (1999) suggests that speakers can also use phonetic information to identify the ethnicity of an unknown speaker. In their study, 50 Caucasian speakers of Standard American English (SAE) were able to correctly identify the ethnicity of twenty speakers[5] of Chicano English (ChE), African American Vernacular English (AAVE) and SAE over seventy percent of the time, with most of the errors coming from mistaking AAVE for SAE. This identification was based on the word *hello* alone, which they took from an original recording of the introductory phrase, *Hello, Im calling about the apartment you have advertised in the paper*. A post-hoc analysis of the stimuli showed that four phonetic features appeared to be significant in distinguishing the dialects: F2 in DRESS, the duration of the initial syllable /hE/, the harmonic to noise ratio (HNR), and where in the word F0 peaked. However, none of these features alone could distinguish all three dialects, and the F2 of DRESS was the only feature that proved significant in distinguishing ChE and AAVE from SAE in a Scheffe test. Furthermore, these phonetic features had not been previously documented as being distinctive phonetic features of the dialects.

Similarly, participants in an experiment by Gordon (1997) were almost in full agreement in assigning a speaker of Broad NZE (as opposed to General or Cultivated NZE) the lowest income, and associating her with a photo of a woman wearing stereotypically lower class clothing[6]. The vehicle for the accent was a letter "specifically designed to elicit phonological variables which have clearly recognizable variants in New Zealand speech" (Gordon 1997, 52), such as TRAP, DRESS, KIT, the merging diphthongs NEAR and SQUARE and /l/ vocalisation. Assumedly, it was the variants used by the speaker that led to how she was perceived by participants, though in having different speakers, this was not strictly controlled for. For example, the pitch, prosody, friendliness or fluency of the speakers may have been the cause of the effect.

This previous work suggests that listeners do use phonetic realisations as cues

[5]Baugh, one of the authors of the paper, also contributed three samples of himself speaking in each accent, all three of which he was well acquainted with.

[6]Creating the outfits that typically summarised working, lower-middle and upper-middle styles was a class project for a year 12 sewing class.

to speaker attribues, but none of the studies have tested the social saliency of a single variable in a systematic way. By the term 'salient', I mean that the realisation of the variable alone has strong enough social associations to alter the perceived social attributes of a speaker. For the main experiment of my thesis, the variable needs to be salient within a whole sentence filled with potentially conflicting speaker information. Thus, in Chapters 3 and 4 I describe two experiments I ran to test the saliency of a number of variables in NZE before I conducted my grammaticality judgements.

2.4 Grammaticality judgements

> "*The fundamental aim in the linguistic analysis of language L is to separate the grammatical sequences which are the sentences of L from the ungrammatical sequences which are not sequences of L and to study the structure of the grammatical sequences... One way to test the adequacy of a grammar... is to determine whether or not the sequences that it generates are actually grammatical, i.e., acceptable to a native speaker, etc*" (Chomsky 1957, 13).

Grammaticality judgements (GJs) have "traditionally been the primary and privileged data for categorical grammatical models" (Bresnan 2005, 17), long used by syntacticians as a means to affirm the (im)possible constructions of a language, and thus to uncover the underlying rules of the grammar. While the empirical evidence that comes from searches of corpora can show the frequency of a construction, GJs offer a way to test the well-formedness of infrequent or potential sentences. They are also the only real source of explicit negative information. At their simplest form, they consist of asking the informant(s) whether a sentence is 'grammatical' or 'acceptable', or for other such intuition-based judgements. More non-direct questions or tasks are also occasionally used, though Heide (2002) argues that the results are less relevant because "it is the experimenter who infers the items evaluation from the subjects performance" (Heide 2002, 94).

If taken as a pure and true representation of a speakers internal grammar, a successful grammar of syntax would parse all sentences judged as grammatical by native speakers but be unable to parse those sentences judged as ungrammatical.

2.4.1 Methodological Issues

"*Should we linguists be worried? I think so*" (Schütze 1996, xi)

For most of the last half century, most GJs in papers had been elicited by the author of the very same paper, consulting his or her intuitions. While this sort of 'couch linguistics is not entirely unreliable in terms of absolutely grammatical and ungrammatical sentences, it becomes insupportable with the less clear cases that much of theory hinges upon. The situation lead Labov to come to "the painfully obvious conclusion... that linguists cannot continue to produce theory and data at the same time" (Labov 1972, 199).

More and more linguists, then, are starting to elicit judgements from a sample of nave speakers, and, in an increasing number, are running their results through statistical analyses to ensure that any perceived patterns are significant. Non-linguist informants unaware of the theoretical implication of their answers are less likely to let these sorts of expectations bias their judgements. Statistical analysis lessens the effect of the random and contaminating factors that behavioral tasks like GJs can involve, and can expose significant patterns that were hidden in the noise of the raw data, or, as in the case of Hirst's (1981) statistical analysis of work done by Ford, Bresnan & Kaplan (1982), can nullify eye-balled patterns as being insignificant.

How judgements are elicited has also become a cause of concern, and the great variation in methodologies is no doubt responsible for the great difference in results across studies. From the instructions, to the scale, to the presentation of stimuli, different methods are used at every step, and could all result in different results between studies, and different interpretations of those results.

In his comprehensive discussion of grammaticality judgements, Schütze emphasises the importance of instructions. Whether using the terms 'grammatical and 'acceptable, he believes it is important to define these terms for participants:

> "*even subjects who are supposedly experts on language can not be expected to know what linguists mean by grammatical... If you do not explain to subjects what you want, each one takes his*

> *or her own interpretation*" (Schütze 1996, 132, emphasis in original).

Of course, the problematic assumption here is that all linguists would agree on what we meant by 'grammatical' ourselves, and this undoubtedly reflects the theoretical framework which we are working in. However, Schütze (1996) also cites a study by Cowart (1993) that suggests if other factors are controlled for, the definition of grammatical is not actually an influential factor in peoples ratings (Schütze 1996, 133).

Instructions do provide a good way to steer participants away from objections based on prescriptivism, which are certainly not what we are after, but are also certainly tied to non-linguist notions of grammaticality. In fact, it may be impossible to separate the two entirely: "it is not yet clear whether we can induce (subjects) to exclude prescriptivist knowledge from their judgements" (Schütze 1996, 83). Schütze's solution appears to be to avoid well-known prescriptivist cases, but (especially if we are testing socially-variable constructions) this is not always possible.

Another important question in designing the methodology of a GJ task is what scale participants should respond on. Binary grammatical and ungrammatical forces binary judgements just as a graded scale can force graded judgements. Most studies use graded scales, but it is often unclear whether (and where) there is a grammaticality threshhold on the scale (Nagata (1988), for example, has only one rating that equates to 'grammatical', and everything below is ungrammatical on a sliding scale). It is also unclear whether participants treat the scales as linear, such that, say, the difference between giving a rating of 1 or 2 is the same in magnitude as the difference between giving a rating of 5 or 6. This had lead some researchers to use Magnitude Estimation Tests, where participants are not given a predefined scale, in eliciting GJs (see Bard, Robertson & Sorace 1996).

Usually stimuli are presented in written form to informants, and a study by Vetter, Volovecky and Howell (1979) showed no effect in responses to stimuli that were presented auditorally or visually. Cowart (1997) suggests that it is best to present stimuli in written form to temper confounding factors that might arise from using speech. However, Kitagawa and Fodor (2006) make a convincing argument that, at least in certain cases, audiovisual presentation

may be best. Looking at sentences that are grammatical with a certain sort of prosody, they found that participants appeared to read the sentences with a default prosody that made them ungrammatical. By presenting the stimuli auditorially, the prosody was in the control of the experimenter, and not the unknown imaginations of the participants.

2.4.2 What do grammaticality judgements mean?

It is undoubtedly not true that GJs are a perfect reflection of the mental grammar. As Bever and Carroll put it "intuitions are not empirical primitives but complex behavioral performances in their own right" (Bever & Carroll 1981, 232). At the same time, it is unlikely, if various contaminating factors are considered, controlled and accounted for, that they offer us no insights into the language faculty. Rather, they provide us with "indexical, that is, causally related, symptomatic evidence for the character of underlying mental representations" (Pateman 1987, 100). The difficulty, then, for linguists is to separate the non-linguistic factors from the linguistic ones. In particular, when is the variation and gradation that consistently shows up in GJs reflective of the grammar, and when is it not? When are influential factors language internal and when are they language external? As Schütze says:

> "*It could be the case that properties such as context dependence and susceptibility to training effects belong to separate modules of the mind that are implicated in judgement behavior but not in other forms of behavior... At another extreme, it could the case that these properties are inherent in the cognitive substance on which language and all other higher cognitive functions are built*". (Schütze 1996, 15).

Unsurprisingly, peoples stance on this tends to depend on the theoretical framework they subscribe to. Some, like Joos, want anything that is uncategorical to be struck from study: "All continuities, all possibilities of infinitesemal gradation, are shoved outside of linguistics in one direction or the other" (Joos 1950, cited by Manning (2003, 290)). Chomsky allowed for three levels of violations (Chomsky 1965). I, like Vogel, believe that if controlled and statistically verified experiments show systematic variation or gradation in grammatical responses "it is very likely that the factor that

caused this intermediate status is grammar-internal. At least, this should be the null assumption" (Vogel 2006, 251). Any other assumption lets the linguist again choose which responses do and do not count on an ad hoc basis, in which case they might as well be using their own intuitions.

In a usage-based model of language, the variation and gradation seen in GJs could be interpreted as reflecting the robustness of the construction in the mind. That is, "(i)ntuitive contrasts in grammaticality that many linguists have reported seem to reflect probabilities rather than categorical constraints" (Bresnan 2005, 1). Crocker & Keller (2006) review some recent literature looking at the degree to which probabilities can be "reinterpreted as degrees of grammaticality" (Crocker & Keller 2006, 240). While the literature they survey suggests there is no direct correlation between probabilities and grammaticality ratings[7], there is enough evidence to suggest that "language experience ... determines (or at least influences) the way speakers make GJs" (Crocker & Keller 2006, 240).

2.5 Aim of this thesis

Speech perception studies in phonology have shown that a signal can be understood differently depending on who listeners believe the speaker is. This supports exemplar models of language use, which say that encountered instances of speech are stored with speaker information, and such information is crucially used in speech processing.

There is a growing amount of evidence that also supports usage-based approaches to morpho-syntax. If ET in particular was used, then we would expect to see speaker effects in the perception of morpho-syntactic constructions analagous to the ones we have see in phonology. If a construction is used more by one group than another, we could expect the perception of that construction to alter depending on which group a listener believes the speaker belongs to.

[7]One of the questions my thesis raises is which factors should be taken into account when calculating the probability of a particular construction, a point I discuss further in 6.2.2.

Grammaticality Judgements are one way that we might uncover such perceptual differences. Though problematic in themselves, if used in a controlled experiment and verified by statistics, it is possible that they tell us about the internal grammar of participants. Thus, if we can alter GJs based on speaker information, this would be a strong indicator that that speaker information was part of the internal grammar of language users.

One possible way to alter speaker information in a tightly controlled manner is to synthetically alter the realizations of socially-salient phonetic variables. In Chapter 3 and 4 I describe experiments that test the social saliency of different variables in NZE. After all, it must first be shown that a variable carries sufficient social meaning to alter perceived speaker attributes before we could ever expect it to alter grammaticality judgements.

In an experiment that I describe in Chapter 5, the most socially salient of these tested variables, phrase final /t/, was manipulated in sentences that participants rated for grammaticality. It is this experiment that I used to test my primary hypothesis: morpho-syntactic constructions are stored with speaker information. The implications of the results are discussed in Chapter 6.

Chapter 3

Pilot Study

The motivation for this pilot experiment is the assumption that the sort of phonetic variables that could alter GJs are those features which are saliently tagged with important information about a speaker's identity. We know that in production there are a number of variants that are used more by certain social groupings than others, based on factors like the speaker's age, class, sex or ethnicity. Furthermore, a small number of perceptual experiments have suggested that listeners do use such variables in making judgements about a speaker's identity (see 2.3).

However, none of the previous studies have explicitly tested whether a particular variant can alter perceived speaker attributes. Furthermore, in the GJ Experiment described in Chapter 5, participants are necessarily presented with full sentences. Therefore we need to know if there are any variables that are powerful enough to alter how the speaker is perceived in a sentence that is full of other, potentially contradictory, speaker information. Additionally, as the participants in the GJ Experiment would be New Zealanders, we needed to test the saliency and social associations of different realisations of particular phonetic variables within NZE.

Seven variable features of NZE were trialled in an experiment where participants were asked to rate the age and social class of speakers based on small extracts of their speech. Participants would hear an extract twice, each with a different realisation of the variable in question, which had been manipulated via synthesis, splicing or cutting. If the variants were socially salient enough, we could expect to get different class and age ratings for the speaker

reflecting the documented distribution of the variants in production. If our manipulation failed to elicit such a response, presumably the different variants would be unable to cause a grammaticality rating change. This would signal the need for either the manipulation of more features or at least the use of more marked features in the speech extracts.

The outline for this chapter is as follows. In section 3.1, I briefly describe seven phonetic variables in NZE, which were tested in this pilot study. Section 3.2 contains the methodology for the experiment, and 3.3 the results of the experiment. A summary of the results is given in 3.3.1 and a discussion of the results in 3.4.

3.1 Variables

Seven candidate phonetic features were selected for the testing based on recent studies of NZE: intrusive /r/, phrase final /t/ (deletion and affrication), DRESS, KIT, and TH-fronting (word-initially and in the pronunciation of the word *with*). These features can all vary in their realisation, and amongst contributing factors to the variation are the social attributes of the speaker. A brief discussion of each of these variables follows.

3.1.1 Intrusive /r/

The term 'intrusive /r/' refers to the phenomenon where, in non-rhotic dialects, a sandhi /r/ is inserted intervocalically after morphemes in which there is no orthographic or 'underlying' /r/. For example, a speaker might produce an [r] between the first two words in the phrase *ma and pa*. This is related to but different from linking /r/, when an [r] is inserted intervocalically in a non-rhotic dialect when there is an orthographically present /r/ (eg. *car and bus*).

Hay and Sudbury (2005) document the growth of this variable in early NZE, showing a rise in the presence of intrusive /r/, and a fall in linking /r/. In a closer examination of the variant in contemporary NZE, Hay and Maclagan (forthcoming) suggested a significance of speaker gender, which has males inserting the /r/ more than females, and showed a very significant effect of social class, with professionals inserting the /r/ less than non-professionals.

Interestingly, they also note that rather than merely being present or absent, there is a gradience of /r/ realisation in such an environment, and it reflects the patterns of its overall distribution, so that not only are non-professionals more likely to insert /r/, but they are also more likely to have stronger /r/ when they do so (characterised by a lower F3).

Intrusive /r/ typically cannot follow high vowels. However, it is beginning to appear after MOUTH (Hay & Maclagan (forthcoming)), due to changes in the diphthong which see its previously high offglide weakened (Woods 1997, Maclagan, Gordon & Lewis 1999). This particular use of intrusive /r/ is significantly more prevalent in males than females and again in non-professionals more than professional speakers. It is also understandably affected by the monopthongisation of the diphthong, and highly diphthongal mouth was less likely to be followed by [r] (Hay & Maclagan (forthcoming)).

For this experiment, only examples of intrusive /r/ following MOUTH were used. The hallmark of an /r/ is a low F3, so for this experiment, the F3 of naturally occuring intrusive /r/ was synthetically raised or lowered to create less and more /r/-like tokens, as will be discussed in more detail in 3.2.2.

3.1.2 Phrase final /t/

A recent study into phrase final /t/ in NZE by Docherty, Hay and Walker (2006) shows that the /t/ are generally realised in one of four ways (Docherty et al. 2006, 378):

- An unreleased plosive almost always realised with accompanying glottalisation and often as a glottal stop
- A canonical /t/ - a sustained voiceless closure followed by a clear release without sustained homorganic frication
- A spirantised /t/ - with no closure gap, and frication during the 'stop' interval
- An affricated /t/ - with a closure gap, and a heavily fricated release

The presence of glottalisation is not restricted to unreleased realisations, and Docherty et al. find it with the released variants as well.

The unreleased variant was by far the most common realisation in their study for all of their speaker groups (which included only younger speakers), accounting for over 70% of the data. It was particularly favoured by males, and non-professionals.

In findings similar to those of Maclagan and Hays work on intrusive /r/, the study found that the release of /t/ could be considered as a graded as opposed to binary factor. They suggest that the fricated variants are markedly salient extensions of the released /t/, and their social distribution reflected the social distribution seen in the released variant. Professionals were most likely to release /t/, and when they did, they were most likely to have fricated realisations thereof.

For this experiment two pairs of phrase final /t/ realisations were tested. The first was an unreleased /t/ vs. a released /t/, both preceeded by glottalisation. The second compared responses to two versions of released /t/: a less fricated /t/ against lengthened frication.The cutting and splicing techniques employed here are discussed in detail in 3.2.2.

3.1.3 DRESS

The DRESS vowel in NZE has raised to the point of now being "the closest of the traditional short front vowels" (Maclagan & Hay 2004, 187). For many speakers, it shares the same acoustic space as FLEECE (Maclagan & Hay 2007, McKenzie 2005) and McKenzie shows that for some speakers it is even higher than FLEECE. These changes are most advanced in the young, professional females and all young, non-professional speakers (Maclagan & Hay 2007).

For this experiment, original tokens of DRESS were synthesised so as to have one raised and slightly fronted version, and one lowered and less front version (see 3.2.2).

3.1.4 KIT

The centralised KIT vowel is one of the most noticeable and studied feature vowels of NZE (Bauer 1992, Bauer 1994, Bell 1997, Bell 1999, Watson, Maclagan & Harrington 1998), with Nicola Woods calling it the sound which

"stamps speakers of NZE" (Woods 2000, 115). The relevance of age in the lowering and centralizing of the vowel is well documented (Batterham 1995, Allan & Starks 2000, Trudgill, Gordon & Lewis 1997), with younger speakers having lower and more central realisations of the vowel. Gender also has an effect, and the KIT of males is less centralised and lowered than that of females (Maclagan 1999, Bell 1997), though of gender and age, the latter appears to be more significant (Mcrobbie-Utasi & Starks 2003).

For this experiment, original tokens of KIT were synthesised so as to have one raised and slightly fronted version, and one lowered and less front version (see 3.2.2)

3.1.5 TH-fronting

TH-fronting refers to the substitution of the labio-dentals /f/ and /v/ for dental fricatives /θ/ and /ð/. It is a common non-standard variant that appears in many dialects of English around the globe (for example, Labov (1972) describes it in Philadelphia; Trudgill (1988) finds it in Cockney). Woods (2003) looked at TH-fronting in NZE amongst 14 young, non-professional speakers in word-initial, intervocalic and word final positions. Most fronting occurred word finally, though this is skewed by the inclusion of *with*, which Maclagan suggests is one of the key words in the spread of TH-fronting (Maclagan 2000, 17). *With* is commonly produced with both voiced and voiceless versions of the labio-dental and dental fricatives.

Wood did not explicitly explore the effects of social factors in TH-fronting, but says that in a pilot auditory analysis of more speakers, "older, especially professional speakers were unlikely to exhibit any TH-fronting in the wordlists" (Woods 2003, 51). Maclagan also states that it is used more "by the younger, non-professional speakers and the older, non-professional males" (Maclagan 2000, 17).

Two different environments for the TH-fronting variables were tested in this experiment: one word initially, and one at the end of the word *with*. Instances of /θ/ were replaced with spliced /f/ taken from elsewhere in the speakers recordings (see 3.2.2).

3.2 Methodology

The experiment consisted of one task, where participants listened to fifty-six sentences, which consisted of twenty-eight sentences that were repeated. For each of the seven variables described above there were four unique sentences in the experiment. The repeated sentences differed only in the realisation of a single variable, which had been manipulated synthetically. Participants were asked to rate the age and social class of the speaker for each token they heard.

3.2.1 Stimuli

The majority of stimuli for the pilot were extracted from the Canterbury Corpus (CC), a body of recorded interviews held at the University of Canterbury. The interviews are recorded by third year students as part of their course work, and begin in 1994 with new recordings added each year. Casual interviews usually last around half an hour or more and are supplemented with the reading of a wordlist (Maclagan & Gordon 1999).

Using the ONZEminer software (Fromont & Hay 2008), time aligned transcripts across various CC speakers were searched for the targeted variables. The focus of the search was in the speech of young, non-professionals. Where possible, a total of four speech utterances were selected for each feature based on audibility, clarity, content and recording quality.

However, of the final twenty-eight speech segments, twelve were taken from archived recordings made for Walker (2005*a*). Three professional males, aged 19, 23 and 50, had been recorded into Soundforge 7.0 reading a variety of sentences aloud into a head mounted microphone. There were two motivations for accessing these recordings for this pilot. Firstly, finding four clean and clear examples of the targeted features to synthesize in the continuous, natural speech of the Canterbury Corpus was not easy. Secondly, it was an opportunity to test if read speech (which would be used in the main experiment) would behave differently than natural speech.

The nature of conversation meant that the appropriate phonological environments found in the CC did not always occur in full, clean sentences. A decision was made that for all of the recordings, the segments that partic-

ipants would be played would be as short as possible while still retaining coherence and context for the manipulated words. Therefore, the segments played to listeners were sometimes only phrases, and pauses were often used as boundaries in choosing where to end or start the segments.

Not every unique speech segment had a unique speaker. Five speakers contributed two segments and one speaker contributed three. So there were a total of 21 individual speakers of the 28 segments.

3.2.2 Manipulation Methods

To create a conservative and an innovative realisation of the variables in question, three different manipulation techniques were employed. The vowels and intrusive /r/ examples were synthesised by altering formant measurements. For the TH-fronting and fricated /t/ examples, splicing was employed, whilst to achieve the unreleased /t/ examples, simple cutting was all that was required.

Synthesis

Formant measurements of the vowels DRESS and KIT, and of the sandhi consonant /r/, were synthesized in Praat using a script designed by Paul Warren at the University of Victoria, New Zealand. The script required the manual selection of four points on the relevant spectrogram, one for the beginning of the transition into the area to be manipulated, one at the end of this transition in, one at the beginning of the transition out of the segment and one at the end of this transition out. The script would alter the formants in the middle section with the specified replacement values and create a smooth transition between this synthesised and the non-synthesised areas.

For the vowels, the original F1 and F2 measurements were taken at the most central and stable part of the vowel. The speech extract was then run through the Praat script twice to create two new versions, classed as innovative and conservative. For DRESS, innovative tokens had lower F1 and higher F2 values than the original, to simulate a higher and fronter vowel, and conservative versions had higher F2 and lower F1 values, to simulate a lower and slightly fronter vowel than the original. For KIT, the innovative versions had higher F1 and lower F2 values, to simulate a lower and backer vowel

Table 3.1: Pilot - Original and synthesised formant values for vowels. Capital letters next to items in the 'word' column indicate different speakers.

Variable	Word	Formant	Original	Conservative	Innovative
DRESS	Jeff	F1	350	450	250
		F2	1849	1800	2000
	getting	F1	443	450	250
		F2	2515	2300	2800
	friends	F1	614	800	500
		F2	2817	2800	2850
	Ben	F1	334	450	250
		F2	2140	2000	2300
KIT	think	F1	602	500	700
		F2	2196	2400	1600
	dip	F1	620	500	700
		F2	1941	2100	1700
	lift(A)	F1	459	350	600
		F2	1366	1650	1000
	lift(G)	F1	434	350	600
		F2	1611	1800	1200

than the original, and vice versa for the conservative versions. The size of the shift in formant values ranged considerably, as different tokens required more or less of a shift to affect an auditorially salient shift in the vowel, while still maintaining a relatively natural sounding signal. The original and synthesised formant values are shown in Table 3.1.

Table 3.2: Pilot - Original and synthesised F3 for intrusive /r/

Words	Original F3	Conservative F3	Innovative F3
how old	2028	2400	1800
now anyway	2900	2000	3800
how old	2360	1800	2500
now and	2020	1800	2400

For intrusive /r/, the script held F1 and F2 constant and only altered F3. Tokens were selected where speakers had originally produced an intrusive /r/. The innovative version of a token was manipulated to have a lower F3 than the original, in order to simulate an even more /r/-like token. The conservative tokens had higher F3 than the original, to simulate an /r/-less token. The original and synthesised formant values are shown in Table 3.2.

Splicing and Cutting

For word-initial and with TH-fronting tokens, the original variant was cut from the speech sample and replaced with either a [θ] (for the conservative version) or a [f] (for the innovative version). Both replacement fricatives were taken from somewhere else in the speakers recording, often the wordlist. For the word-initial TH-fronting tokens, instances were selected in which the speaker originaly used a [θ]. *With*, however, barely ever occurred with a [θ]. An attempt was made to take the substitutions from the same environment: word-initial substitutes came from word initial positions, word final from word final. However, this caused some difficulty for *with*, as it was very hard to find word-final [θ] in the recordings of the speakers used. So for two of the speakers, the spliced [θ] variant came from word initial position. An effort was also made to make the contrasting spliced consonants of a similar length.

To create released and unreleased /t/ variants, tokens of phrase final /t/ were found that had glottalisation followed by a /t/ release (which was usually somewhat spirantised). When the release was cut off the end of the word, the glottalisation alone was a sufficient realisation of the /t/. So for this feature, one of the segments was altered by having the release cut off, and contrasted with the unaltered original, where the /t/ was still released.

To create two released phrase final /t/ variants with different levels of affrication, tokens were selected that had reasonable levels of frication following a release. A middle section of the energy was selected and either cut from the consonant or copied and pasted immediately after. This resulted in two variants, one where the original frication had been lessened, the other where it had been lengthened.

3.2.3 Experiment Design

There were two manipulated versions of each of the original 28 selected speech segments, so participants listened to a total of 56 utterances during the experiment. The 'same' segments were usually 30 utterances apart, but range in distance from 10 to 44. The same speaker was kept at least ten utterances apart. To account for the influence that order could have on the results, participants were put into one of two trials. The order of utterances was exactly the same in both trials, but the variants were inverted, such that if the conservative realisation was played first in one trial, it was played second in the other.

The recordings were organised into a Praat script, which was then recorded from computer on to cassette tape and played to participants over headphones on a Sony TCM-5000EV portable cassette-recorder. Occasionally two participants would do the experiment at the same time, in which case a double adapter would be used and they would still have their own set of headphones. Participants were given an answer booklet in which to mark their responses.

Through an instruction sheet, participants were asked to rate the age and social class of a range of speakers. It was stressed in the instructions that they should focus not on what the person said but how they sounded. For the age of the speaker they could circle one of nine options, each of which covered five year increments, as seen in (1).

1. ... Jeff isnt a bad fellow...
   Age: 15-20 21-25 26-30 31-35 36-40 41-45 46-50 51-55 56+
   Class: Working Lower-Middle Upper-Middle Upper

For the class of the speaker, participants were given the four options of working, lower middle, upper middle, and upper (1). These titles were fairly arbitrary, as there is not an openly conceded social class structure in New Zealand, and my participants sometimes expressed surprise to be faced with the labelled categories. As Elizabeth Gordon said in the third of her Macmillan Brown Lectures on New Zealand English:

> "*And it is uncomfortable talking about social class. So uncomfortable that some New Zealanders will say that there are no social class differences in New Zealand. But the linguistic evidence of social class in New Zealand is incontrovertible*".

Some sort of class structure exists, but defining the groups is problematic. Class is related to wealth and education, but not confined to them. It is standard, when trying to ascertain the social class a participant sees someone as belonging to, to ask them what they think the occupation and education of the speaker is. However, in this quick task, such questions seemed onerous. It was decided to leave class undefined to target the general meaning of the word, and as comparisons were primarily intra-speaker, any inconsistencies between participants would have less of an impact.

As seen in (1), the written form of what the speaker was saying appeared next to the token number. As the segments were taken out of context, were often only parts of a sentence, and were extracted from quick and informal speech, this seemed necessary to help participants understand what was actually being said, and thereby be more sensitive to what variants were being used. There were only six tokens in each page of the response booklet, and none of the same sentences appeared written on the same page, nor did participants have enough time to flip back and check their previous answer when faced with an utterance they had already seen.

There was a five second pause after each segment in which speakers circled the age and social class they thought best matched the speaker. This was intentionally brief to capture their immediate instincts. The whole task took just under ten minutes.

3.2.4 Participants

27 people took part in the experiment. They range in age from 18 to 60, and their distribution across the two trials is shown below in Table 3.3. Participants were all given a numerical value reflecting their social class. These numbers were calculated by assigning their parents a New Zealand Socioeconomic Index (NZSEI), as developed by the New Zealand Standard Classifications of Occupations (Davis et al. 1996, Davis et al. 2003). The indices range from 0 to 100, and jobs that receive more income and carry more

Table 3.3: Sex, age and class of participants in the Pilot Study.

Trial	A	B
Total Participants	14	13
Total Females	5	6
Total Males	9	7
Min Age	20	18
Max Age	60	53
Median Age	24.5	24
Min Social Class Index	74	54
Max Social Class Index	157	150

prestige have higher scores on the index. A participant's Social Class Index was attained by adding the NZSEI of their mother and fathers occupations. It is standard practice in the ONZE Lab to assign young students or participants a class score based on their parents occupations instead of their own. This is because, as many young people used in our research are students or in temporary jobs that wont reflect their future career paths, it seems specious to allocate them scores on their current occupations. Furthermore, parents occupations frame the environment under which participants were raised. Unlike standard ONZE practice, I decided to allocate all participants class indices based on their parents occupations, including older participants, so that the scores would be more comparable.

3.3 Results

The ratings from the response sheet were manually copied into an Excel spreadsheet. A linear regression model was hand fit to the data using Harrells Design Library (2004) in R.2.0.0 (Team 2004).The dependent factor, here either the ratings for age (EAGE) or social class (ECLASS), was modelled to see if, how and to what extent it could be predicted by a number of independent effects, which came from information about the speakers, participants, the stimuli and the experiment. The age (SAGE - a binary young(below 30) or old(over 40) variable) and gender (SGENDER) of the speaker were considered, as well as the source/mode of the recording (MODE) and the age in years (AGE), social class (CLASSNUM) and sex (SEX) of the participant. From the stimuli we took what phonetic variable (i.e. intrusive- /r/) had

been manipulated in the sentence (VARIABLE), and whether the realisation of this variable was the innovative or conservative version (REALISATION). Issues relating to the experiment were which trial group the participants were in (TRIAL), the number of words in an utterance (NUMWORDS) and how far into the experiment the utterance was (ORDER).

Table 3.4 shows the factors that proved significant in the final ANOVA for the age ratings. Something is deemed significant if it has a p-value of less than 0.05, and the lower the p-value, the more robust the effect. Table 3.5 shows the coefficients of the model (N=1506, R^2=0.337), which tell us direction and size of the effects. The intercept is the models prediction of the age rating a speaker would receive with default independent factors. For continuous factors, such as ECLASS, the default would be 0. For categorical factors, like the type of phonetic variable that was being manipulated (VARIABLE), the default is the coefficient that comes first in the alphabet, which in this case is DRESS (so DRESS, which does not appear in the table, has a coefficient of 0 by definition).

Table 3.4: Pilot - ANOVA for the age ratings of speakers.

Factor	d.f.	Partial SS	MS	F	P
ECLASS	1.00	250.427	250.427	167.16	<0.0001
AGE	1.00	17.83188	17.83188	11.9	0.0006
SAGE	1.00	97.91332	97.91332	65.36	<0.0001
VARIABLE	6.00	71.2876	11.88127	7.93	<0.0001
MODE	1.00	14.43443	14.43443	9.63	0.0019
ORDER	1.00	14.05063	14.05063	9.38	0.0022
SGENDER	1.00	225.1824	225.1824	150.31	<0.0001
NUMWORDS	1.00	27.43278	27.43278	18.31	<0.0001
regression	13.00	1163.125	89.47116	597.72	<0.0001
error	1492 2235.205	1.498127			

Importantly for the purpose of the study, Table 3.4 shows that there was no overall effect of whether the token had contained an innovative or conservative realisation of the variable, because the factor REALISATION is not in the model, meaning it did not reach significance (p = >0.05). There were other significant factors however. The age rating is a significant predictor of the social class rating, so that the higher one is, the higher the other. Young

Table 3.5: Pilot - Coefficient table for the age ratings of speakers.

Coefficients	Value	Std.Error	t	p-value
Intercept	1.999986	0.246486	8.11399	<0.0001
ECLASS	0.525377	0.040635	12.92904	0.0000
AGE	-0.0104	0.003014	-3.45004	0.0006
SAGE=young	-1.17383	0.145198	-8.08438	<0.0001
VARIABLE=fric /t/	0.002323	0.124113	0.01871	0.9851
VARIABLE=intrusive /r/	-0.03235	0.135052	-0.23952	0.8107
VARIABLE=kit	-0.1839	0.125848	-1.4613	0.1441
VARIABLE=no /t/	0.272562	0.126889	2.14803	0.0319
VARIABLE=TH-fronting	-0.38456	0.137443	-2.79799	0.0052
VARIABLE=with	0.299064	0.14253	2.09826	0.0361
MODE=read	-0.38782	0.124941	-3.10403	0.0019
ORDER	0.006282	0.002051	3.06248	0.0022
SGENDER=m	1.100478	0.089761	12.26007	0.0000
NUMWORDS	0.067699	0.015821	4.27918	<0.0001

participants were more likely to rate speakers as older than were older participants. Older and male speakers were rated as being older. MODE also had an effect, such that speakers from the conversational Canterbury Corpus were generally rated as older than the speakers reading passages. Tokens that appeared further into the experiment and tokens which contained more words also elicited older judgements.

There was also an effect of the type of variable being tested, as can be seen in Figure 3.1. With and release/no-release /t/ sentences were given the highest ratings, whilst tokens with TH-fronting and KIT were given the lowest. While the different sets of sentences were not matched for speaker attributes like age or gender, this effect is significant above and beyond the effect that those factors might have in the model, because the model takes them into account.

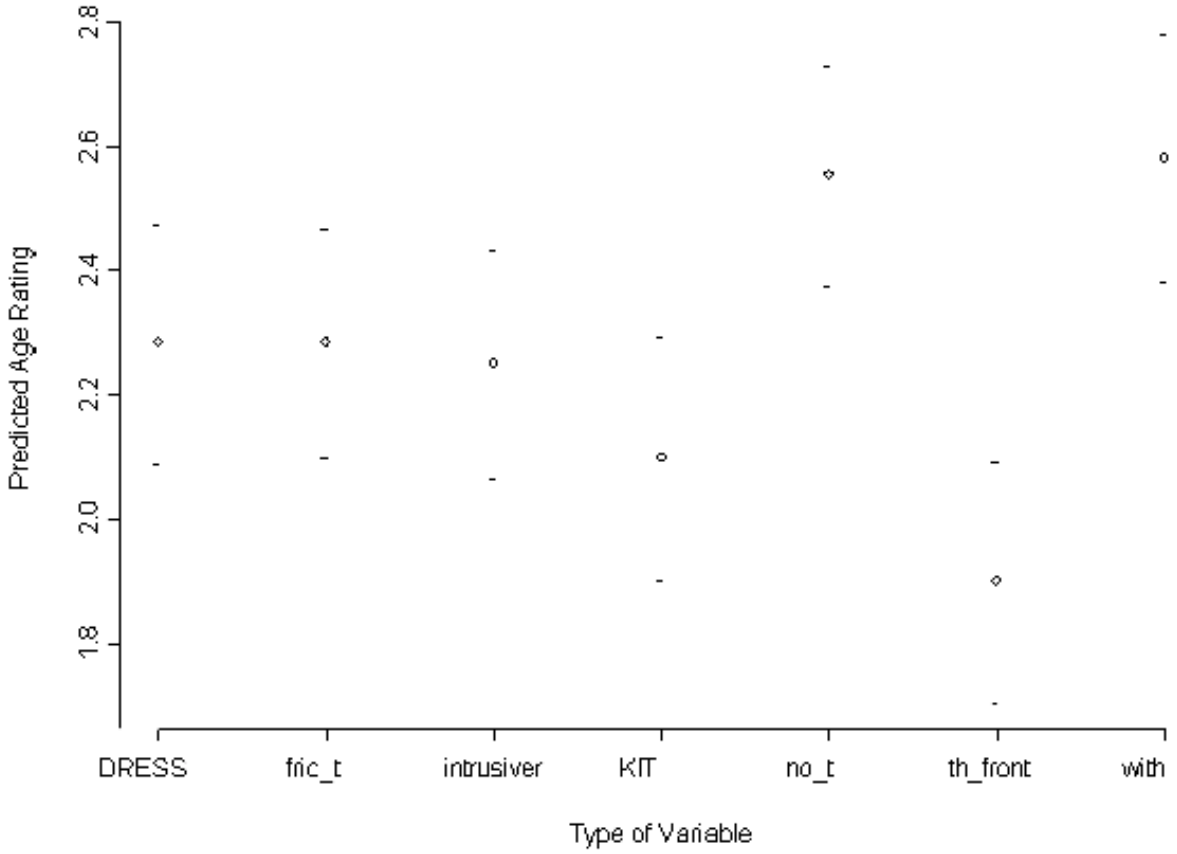

Figure 3.1: Pilot - Models Prediction of effect of type of variable on age ratings.

The significant factors in the social class ratings given to speakers are shown in Tables 3.6 and 3.7 (N=1506, R^2=0.2346). We again see the positive relationship between the class rating and age rating given to a speaker. Other patterns were that older and male participants were more likely to give a higher social class rating to a token, and older and female speakers received higher ratings. The mode of the recordings was again important, but for social class, higher ratings were given to the read passages over the natural speech extracts. Again, the realisation of a variable had no significant effect on the overall class ratings given to tokens.

Table 3.6: Pilot - ANOVA for the class ratings of speakers

Factor	d.f.	Partial SS	MS	F	P
EAGE	1	96.1761	96.1761	177.2	<.0001
AGE	1	4.639973	4.639973	8.55	0.0035
SAGE	1	2.826673	2.826673	5.21	0.0226
VARIABLE	6	79.96679	13.3278	24.56	<.0001
MODE	1	13.85575	13.85575	25.53	<.0001
GENDER	1	6.157875	6.157875	11.35	0.0008
SGENDER	1	3.128521	3.128521	5.76	0.0165
regression	12	256.8782	21.40651	39.44	<.0001
error	1493	810.3363	0.542757		

The type of varable was again important, as displayed in Figure 3.2. However, the effect was different than that seen on the age ratings. Fricated /t/ tokens were more likely to be given a higher class rating, as were the dress and TH-fronting tokens. The lowest class ratings were given to the with sentences.

To further examine any possible effect REALISATION might be having, the models were rerun with the independent effect VARIABLE replaced with a new combined factor, TYPEREAL. TYPEREAL split the seven types of variables into their conservative and innovative realisations, so that where we had 7 possible entries for VARIABLE (i.e., intrusive-/r/), we now had 14 possible entries (i.e., intrusive-/r/-with-conservative). TYPEREAL, like realisation, is not significant in the model, but I include it here as a diagnostic tool so we may see how the realisations patterned in regard to each phonetic variable. In Figure 3.3, which shows the age ratings, we see that for the dress, no-/t/ and intrusive-/r/ tokens, there seems to be a difference between the conservative (c) and innovative (i) variables, that goes in the expected direction so that conservative tokens receive a higher age rating than the innovative tokens. The other types receive relatively similar ratings for either realisation, and it is only in with tokens that there appears to be a difference that goes in the unexpected direction, so that innovative forms receive a higher rating than conservative ones.

Table 3.7: Pilot - Coefficient table for the class ratings of speakers

Coefficients	Value	Std.Error	t	p-value
Intercept 1.760393 0.134763 13.0629 0.0000				
EAGE	0.194347	0.0146	13.3116	0.0000
AGE	0.005311	0.001816	2.9238	0.0035
SAGE=young	-0.20302	0.088961	-2.2821	0.0226
VARIABLE=fric /t/	0.282044	0.073729	3.8254	0.0001
VARIABLE=intrusive /r/	-0.24757	0.07978	-3.1032	0.0020
VARIABLE=kit	-0.3723	0.074515	-4.9964	<0.0001
VARIABLE=no /t/	-0.32631	0.072623	-4.4932	<0.0001
VARIABLE=TH-fronting	-0.01329	0.079999	-0.1662	0.8680
VARIABLE=with	-0.46675	0.082751	-5.6403	<0.0001
MODE=read	0.366202	0.072478	5.0526	<0.0001
GENDER=m	0.130066	0.038615	3.3683	0.0008
SGENDER=m	-0.13498	0.056222	-2.4009	0.0165

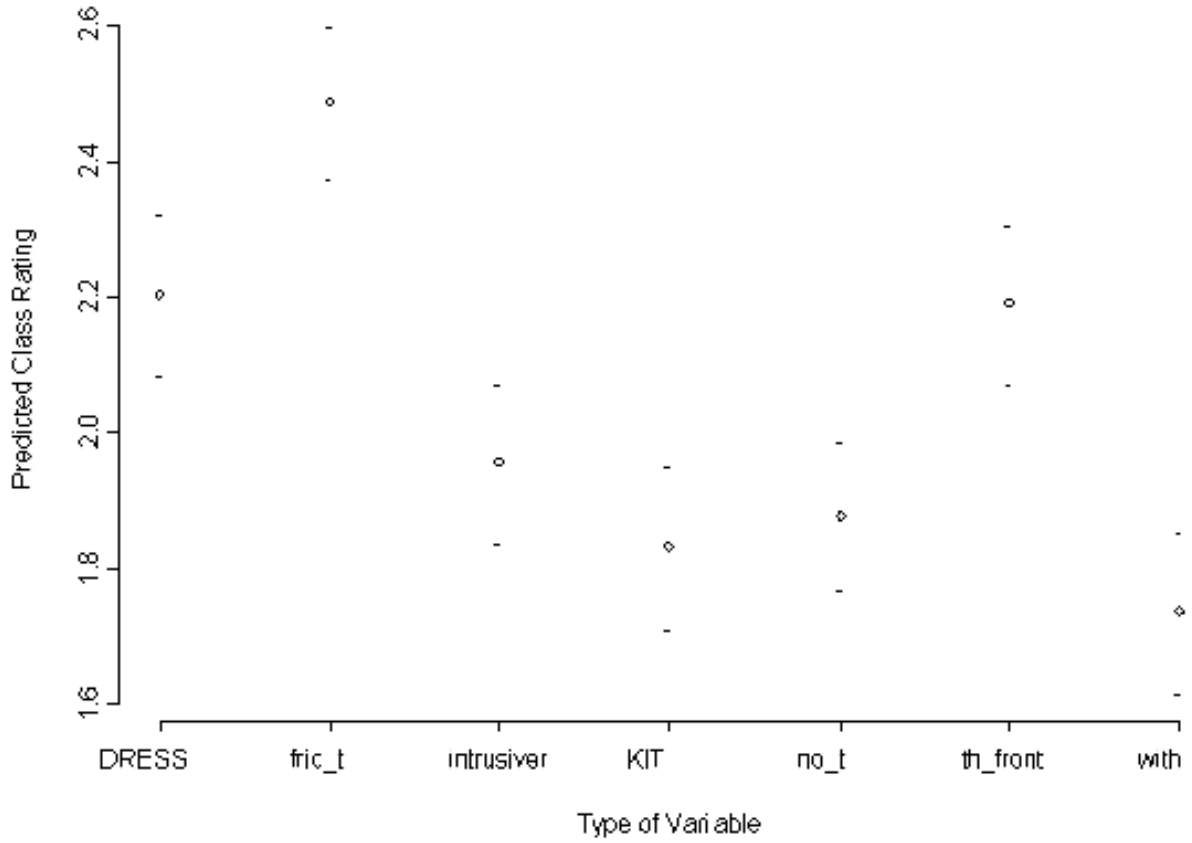

Figure 3.2: Pilot - Models Prediction of effect of type of variable on class ratings.

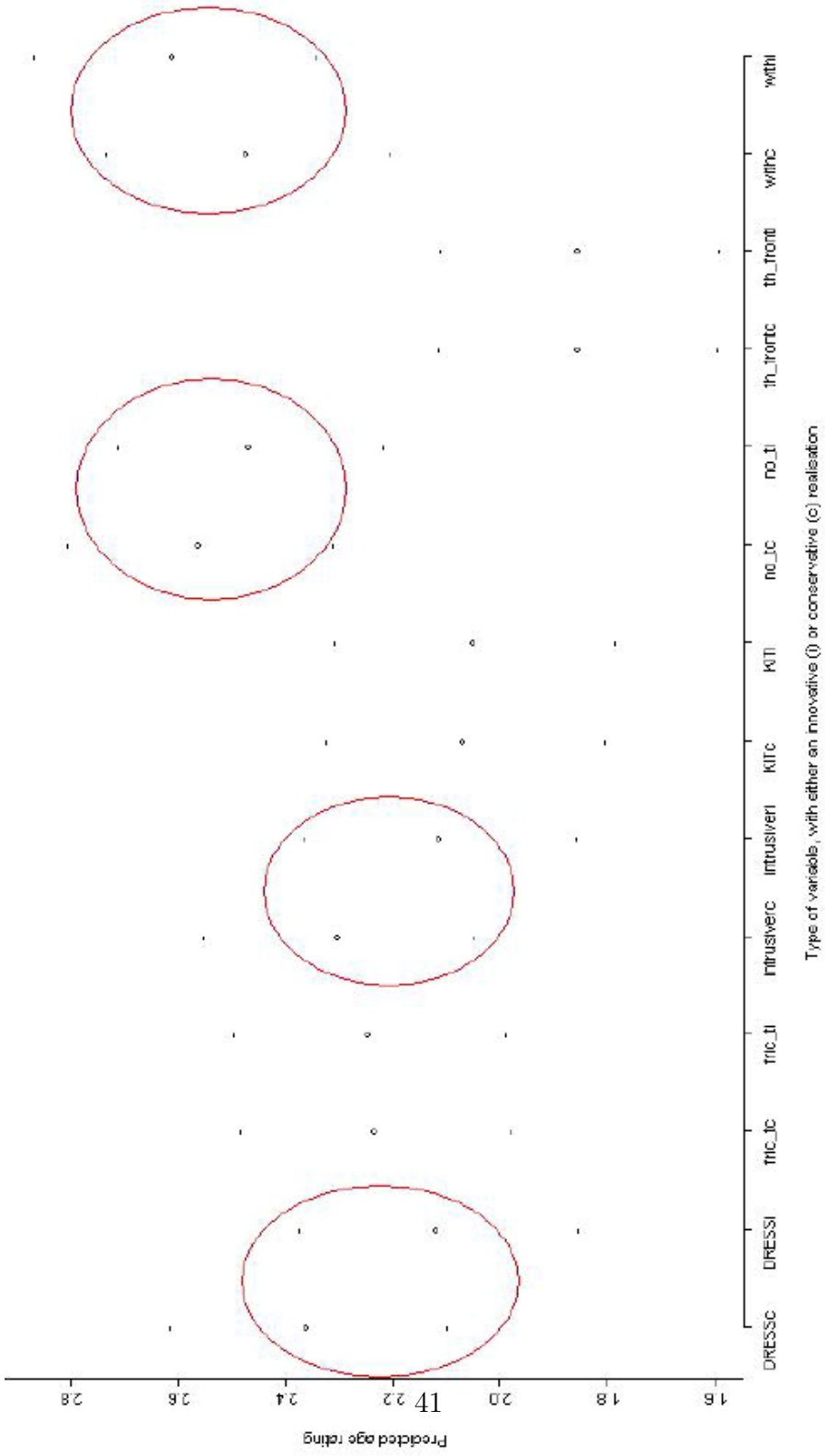

Figure 3.3: Pilot - Models Prediction of effect of type of variable and the realisation of that variable on age ratings. Areas of interest highlighted

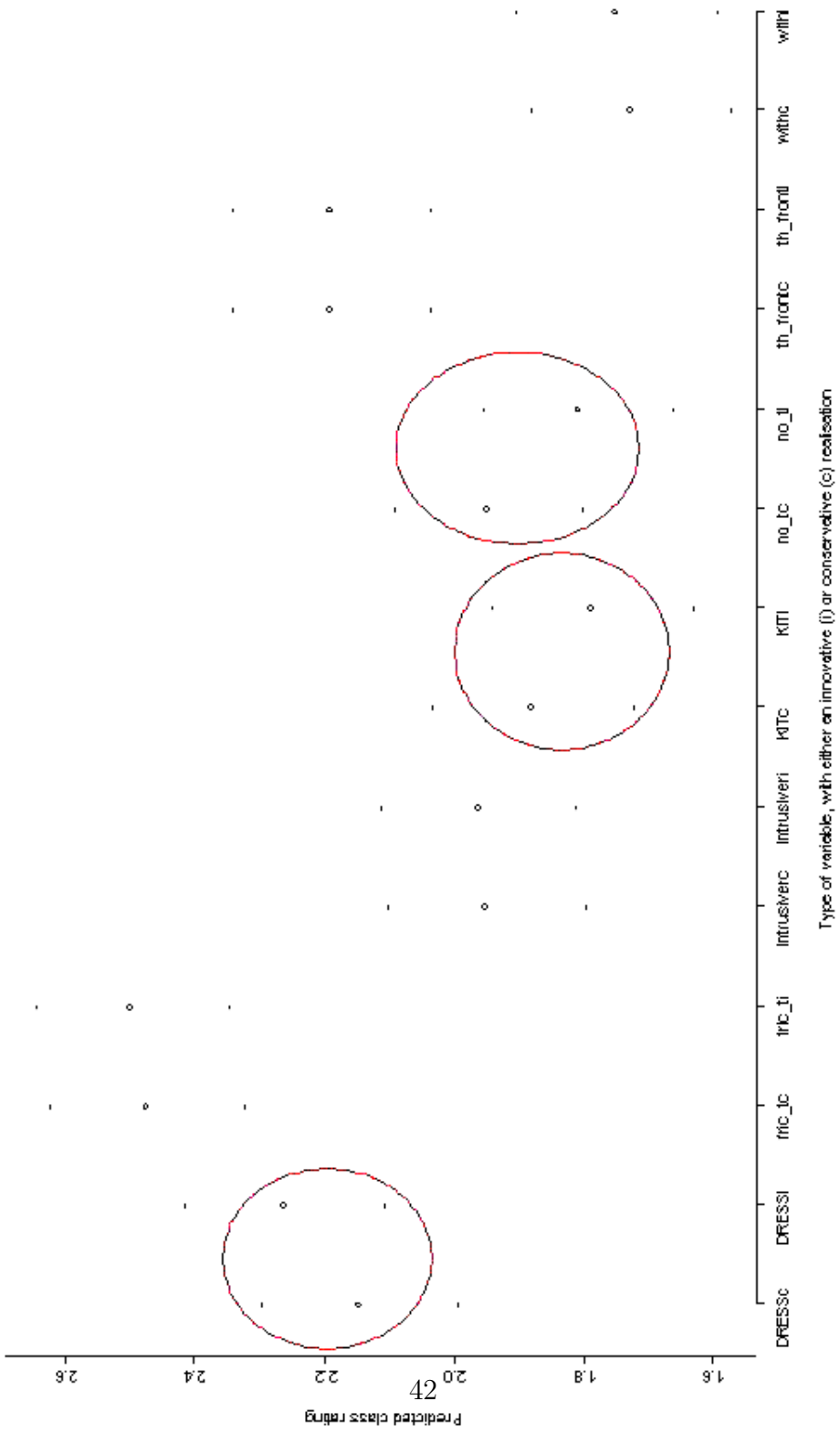

Figure 3.4: Pilot - Models Prediction of effect of type of variable and the realisation of that variable on age ratings. Areas of interest highlighted

For the class ratings (Figure 3.4), there is only a visible difference in the dress, no-/t/ and kit ratings. However, while KIT and no-/t/ pattern as expected, such that the conservative variant receives higher ratings than the innovative variant, in this graph, innovative DRESS variants receive higher ratings than the conservative forms.

A series of Wilcoxon matched-pair tests were run to compare the average responses to the conservative and innovative realisations for each variable under consideration. There was an almost significant difference between the responses for the age ratings of intrusive /r/ tokens (p=0.055) and a just significant rating for the difference between the responses for the class ratings of no /t/ sentences (p=0.049). In both instances, the conservative realisation received the higher rating. There were no other significant results for the other types of tokens.

3.3.1 Summary of Results

There were a variety of factors that were not seriously controlled for in this pilot study, so it is wise to consider the results as reflecting potential areas of interest, as opposed to providing conclusive evidence. The age and class ratings given to a token were highly correlated, such that the higher one, the higher the other. This is an effect that has been found before (Drager 2005). Unsurprisingly, older speakers received the highest age and class ratings. Younger participants rated speakers as older but of a lower social class than did older participants. For the age ratings, participants rated men as older and women as having a higher social class. Male participants rated speakers as younger but women as having a higher social class, which may be related to a gender difference in class ratings that Gordon (1997) found, where male participants were more reluctant to give out low class ratings, and even gave out higher ratings to participants who they perceived as lower class, but felt sorry for.

For the age ratings, there were effects of when in the experiment a token was heard by participants and the number of words in a token. As the experiment progressed or the utterance got longer, age ratings became higher. In both the age and class ratings, there was an effect of the mode/source of the stimuli, though it went in divergent directions. The age ratings were highest for the natural speech extracts that came from the CC, but the class ratings

were highest for the read tokens that were recorded for Walker (2005*b*).

There was an effect of the type of variable that was being manipulated in both the age and class ratings, though again, it worked somewhat differently in each. The with and no-/t/ sentences received the highest age ratings, whilst the TH-fronting sentences received the lowest. However, the *with* sentences received the lowest and the TH-fronting sentences received close to the highest class ratings. There were so many differences between the sentences it is difficult to draw many conclusions from this effect. It is worth noting, however, that the fricated-/t/ sentences received the highest class ratings convincingly. Both the 'conservative' and 'innovative' variants of these particular sentences were released /t/, just with varying degrees of frication. This could suggest, combined with the other patterns in the data, that the released /t/ is a particularly salient marker of professional speech.

The realisation of variables had no significant effect in the linear regression models of either the age or social class ratings. In the patterning of the TYPEREAL graphs (Figures 3 and 4), it seems that for the KIT, DRESS, intrusive /r/ and especially the no-/t/ manipulations, there may be a difference in ratings between the two realisations, such that the conservative variants are rated higher than innovative ones. In Wilcoxon tests, this difference is almost significant for the age ratings of intrusive /r/ and just significant for the class ratings of the no-/t/ tokens.

3.4 Discussion

The different realisations of the phonetic variables appeared to have no overall effect on the age or class ratings, but a closer look at the data revealed that the manipulation of the no-/t/ variables was most successful, and also that there might be reasons to further explore KIT, DRESS and intrusive /r/. There are a number of reasons why a variable might be more effective at altering perceived speaker attributes than others, some to do with the manipulation methods, and others to do with the variables themselves.

The way in which the two variants were created was more successful for some variables than others. The synthesis of the vowels was probably simply not good enough. Though the resynthesised vowels sound fairly natural, there

was usually some remnant of the synthesis process in the sample, often as a sort of click. This may have distracted participants from hearing the vowel clearly. Despite splicing in distinct [f] and [θ] into the TH-fronting examples, the resulting segments generally did not sound very different. This is surprising given the huge stigma attached to TH-fronting in NZE, and suggests that perhaps there was some remnant of the original consonant in the surrounding vowels that rendered the conservative and innovative versions non-distinct.

Phrase final /t/ might also be more auditorially salient than the other variables examined. A released /t/, especially with frication, is quite turbulent, and the comparison with an unreleased /t/ is marked. That is also perhaps why the lengthening to the frication didnt work, because the most socially meaningful difference is in the presence and absence of the release, and not in the length of the release. The /t/ was also produced phrase finally, so it was the last phoneme participants heard before making the ratings.

3.5 Conclusion

There was not an overall significant effect of the realisation type on the age and social class ratings of speakers in this experiment. However, as the results were examined in more detail, it appeared that the effect might be working in phrase final /t/. The next chapter describes a more controlled experiment that tested the social-saliency of phrase final /t/ more thoroughly, in the same sentences that would be used in the GJ experiment.

Chapter 4

Speaker Perception Experiment

The results of the pilot experiment described in the previous chapter suggested that the manipulation of phrase final /t/ within tokens of speech may be able to alter the perceived age and social class rating of the speaker. This chapter describes an experiment that more thoroughly investigated the social saliency of phrase final /t/ , this time within the same recordings that were to be used in the GJ Experiment described in the following chapter.

This experiment was designed in conjunction with the main experiment for two reasons. Firstly, it was deemed wise to test the social saliency of phrase-final /t/ in the same sentences that would then be rated for grammaticality in the GJ task. If the /t/ worked in this Speaker Perception (SP) task and not the GJ task it would potentially tell us little if different sentences were used in the two tasks. Furthermore, the results from the SP task could be included in the model of the GJ task as potential predictors of the grammaticality ratings. Secondly, it was more time efficient to record only one set of sentences.

It should be noted that the experiment described in this chapter was done after another very similar experiment (Walker 2007) that is not described in this thesis due to overlap. This experiment had tested both intrusive-/r/ and phrase final /t/, but only in the target constructions of the GJ experiment (this experiment importantly tested the filler sentences as well). The results suggested that while phrase final /t/ could significantly alter the age and class ratings of the speakers, intrusive-/r/ did not (though sentences which had both manipulated intrusive /r/ and phrase final /t/ were the most affected).

The findings of the experiment were presented at the 11th International Conference of Phonetic Sciences in Saarbrucken, and the resulting paper can be read in the proceedings.

4.1 Methodology

The experiment consisted of a single task where participants were asked to estimate the age and social class of a speaker based on a single sentence. They would hear each sentence twice throughout the experiment, and the two presentations would differ only in the realisation of the final /t/. Unlike in the pilot, stimuli for this experiment were designed and recorded specifically for the thesis.

4.1.1 Sentences

It was decided that the sentences manipulated in the experiment would be ones we expected to use in the main experiment. This was partly to save time in recording new sentences later, but also served as an opportunity to test how the variables would be rated in the constructions we were interested in. Furthermore, if the same sentences were used, we would be able to put the results from this experiment into the statistical model for the GJ experiment as independent effects.

The aim of the main experiment is to test the effect of varying phonetic detail on the grammaticality ratings of non-standard morphosyntactic constructions in NZE. For this study then, the target sentences needed to contain constructions that show social variation in their distribution in NZE. Non-standard preterite forms and the use of HAVE-*got* to denote possession were the two constructions selected, and are described below. Some control sentences were also created and recorded for the GJ task. These control items tested if there might also be an effect of /t/ realisation on sentences without obvious socially marked alternatives, either because they were either so standard (the NORMAL sentences), or contained non-native style mistakes that are not documented at all in production (the BAD sentences).

Target Sentences

Non standard Preterites

Tagliamonte (2001) calls the alternation of *come* and *came* in preterite contexts "one of the most familiar nonstandard features of English dialects" (Tagliamonte 2001, 42) . This non-standard variant is present in NZE, and has been studied in conjunction with other participle forms that can appear in past-reference contexts, such as done and seen (Quinn 1995, Quinn 2000, Durkin 1972).

The biggest factor in the production and acceptance of the form appears to be the social class of speakers, with professionals less likely to accept or use it, though gender may also play a more minor role (Quinn 1995). Most studies have focused on young speakers only, but some recent work by a second year sociolinguistics class at the University of Canterbury shows that age does have a role, with younger speakers using the form *seen* more than any other group (Heidi Quinn, personal communication). The frequency and distribution of the non standard form is lexically specific, and Durkin (1972) had evidence that *done* was more common than *seen* and Quinn (1995) that *come* was accepted more than *seen*. The recent University of Canterbury study suggests that done was barely used at all by speakers in the Canterbury Corpus (see 3.2.1), though *come* and seen were fairly common. I also examined preterite forms in Walker (2005*b*), and found that *come* and *done* were not only rated the least grammatical of all the non-standard preterite forms, but that they also garnered the most significant differences in grammaticality ratings of the older and younger speaker's sentences. It was decided, for this study, to focus only on *come* and done.

Tagliamontes examination of the *come/came* alternation in Yorkshire English found that preterite come occurred most frequently with first and third person singular subjects. She also discovered, in her older speakers at least, that *come* was most likely to occur with a verbal particle, such as *up, over, out* etc., which she suggests might mean the difference between using *come* and *came* is one of verbal aspect. Therefore the *come* sentences in this study were all accompanied by a particle and all had a third person singular subject.

Below are the ten COME sentences (1-10).

1. George come over last night.
2. She come back into the house and forgot about it.
3. A lady come through last night.
4. Tom come back while Susan was walking out.
5. We were ready quite early, but she come round really late.
6. Eventually she come back and we worked it out.
7. I called her name and she come down from the hut.
8. When she come in I thought she looked great.
9. George come over in his new car and took me out to eat.
10. He come up from the basement and talked to Brett.

Because come and done can occur standardly in perfective and pluperfect constructions after HAVE, care was taken to avoid the onset /d/ of *done* being heard as a contracted had, even though in most of the sentences a pluperfect reading would be unnatural. Therefore all words preceeding *done* ended in a consonant, except for (17) and (20), and in both of these sentences the pluperfect reading would be unlikely. Furthermore, all subjects were kept in third person singular, where HAVE-deletion is documented as barely occuring (Quinn 1995, Holmes, Bell & Boyce 1991). Below are the ten DONE sentences (11-20).

11. I just can't forget what Greg done to Matt.
12. George done the dishes late last night.
13. I know what Liz done but he deserved it.
14. Pam done it first thing that morning before eight.
15. I've often wondered just how Frank done that.
16. Well, Jen done it, but much too late.

17. He only done it because I was upset.

18. Ben done it yesterday and I guess it looks alright.

19. Hamish done it before the fight.

20. Rachel talked to me before she done it.

Possessive HAVE-*got*

The use of possessive HAVE-*got* has increased in both American and British varieties of English (Noble 1985, Tagliamonte 2003). Quinn (2004) compared the use of possessive HAVE-*got* in positive present tense sentences across three corpora of NZE and found that speakers from the youngest corpus (CC) used the form significantly more than those from the older corpora. She also found interactions with the social class of speakers, as well as their gender. Non-professional males were more likely to use the form than their female counterparts, whereas professional females used it more than male professionals. Splitting the CC into older and younger speakers, she found that for the older speakers, men used the construction more, but for younger speakers, females used it more.

For the HAVE-*got* sentences in this experiment, all the subjects were third person singular, and so all the realisations of HAVE were contracted has. The sentences were designed to end with the HAVE-*got* phrase, so the actual construction in question was also undergoing the phonetic manipulation (21-30).

21. Now I think thats all she's got.

22. How about selling the one Adam's got?

23. I asked for more but he said thats all he's got.

24. I wonder how much money he's got.

25. I bought her something she's already got.

26. That farm is all she's got.

27. I hope that isnt the only plan he's got.

28. Shed give away everything she's got.

29. Its the only home Jane's got.

30. He was asking how much cattle Henry's got.

Control Sentences

In addition to the Target sentences, participants also rated the age and class ratings of 20 control sentences. The NORMAL sentences are standardly grammatical and had no obvious production biases to any particular social groups. The BAD sentences contained errors that are not documented as occuring in NZE, and as such, should also have no obvious production biases to any particular social group because they are not used by any groups. The hypothesis is that the manipulation of the phrase final /t/ should have no effect on the grammaticality rating of these control sentences, if the ratings do indeed reflect experience. However, for this SP Experiment, we have no reason to believe that manipulation of the /t/ would be any more or less effective in altering the age and class ratings in the Controls than in the Targets.

Normal Sentences

There was little stipulation in the design of the NORMAL sentences except that they ended in phrase final /t/ and did not contain constructions that are marked in any way. They are given below in (31-40). As a caveat, however, it should be pointed out that the author is a young female, and as these sentences have not been pre-tested, it is still possible that something in the wording is unintentionally youthful.

31. Geoff has been trying hard to get our vote.

32. The door slammed and it gave me such a fright.

33. I'm so hungry but theres nothing to eat.

34. My sister named our dog and I named our cat.

35. Last week the boys were involved in another fight.

36. John sent me the link to his new site.

37. He always makes my coffee too sweet.

38. Dave yelled at her and she was really upset.

39. I've decided on my dress but I cant find the right hat.

40. I think we were all really affected by it.

Bad Sentences

The ten BAD sentences contained constructions that were not documented as appearing in NZE at all. However, though they might be unnatural and ungrammatical, they were designed to be understandable. Four of the sentences had pronouns with incorrect case, so that, for example, a third person feminine singular subject would be her and a first person plural object would be we (41-4). Three of the sentences had had plural subjects with verbs marked for singular subjects (45-7). The last three sentences were verb final (48-50).

41. Us left the club really late.

42. Him jumped out and gave we such a fright.

43. I think her is out on a date.

44. If you give it to we, us can fix it.

45. Beth and Freddie has a really friendly cat.

46. Jason and Laura was talking really late into the night.

47. We was really feeling the heat.

48. Alice with her older sisters would always fight.

49. The car the corner at a frightful speed hit.

50. The dress her much better than me fit.

4.1.2 Speakers

Five young females (Table 4.1) with theatrical experience[1]were recruited and recorded into Sony Sound Forge via a head mounted microphone connected to a USB pre. The women were all native speakers of NZE, all were tertiary educated and they ranged in age from 19 years of age to 27. Their social class indices were attained as explained in Section 3.2.4, though the markedly lower scores of Speakers L and K highlight some of the problems with using a standardised class calculator like the NZSEI, as I would not rate either of them as belonging to a significantly lower class than the others[2]

All women read all of the sentences, and were told to make them sound as natural and conversational as possible. At the same time, they were coached into consistently producing /t/ with a release phrase finally.

Table 4.1: Age and class of the five women recorded for the Speaker Perception and GJ experiments..

Speaker	Age	Social Class Index
A	25	115
E	19	134
K	28	65
L	26	44
S	20	102

[1]The theatrical experience means that they read the sentences naturally and comfortably, but potentially also makes them speak more standardly than might be expected from their age and class. Speakers S, L and K have been theatrically trained.

[2]The markers of this thesis were both concerned by the social class indices of speakers L and K, especially as these speakers were rated as being generally of a higher class than the other speakers in the Speaker Perception Experiment. I agree, and in hindsight should have dropped the NZSEI entirely for discussion of these speakers, as it incorrectly made what was really quite a homogenous group seem markedly mixed

4.1.3 Manipulation

For each different sentence type (i.e., a COME or a HAVE-*got* sentence), two unique sentences were taken from the recordings of each female. Sentences were chosen where the speaker had produced a natural sounding phrase final /t/ which exhibited glottalisation followed by a release. The manipulation of the /t/ was achieved with the same simple techniques described in the pilot (3.2.4): the release would simply be cut from the end of the sentence, so that the original version was used as the conservative form and the cut one, with only glottalisation to indicate the /t/, was used as the innovative realisation.

4.1.4 Experiment Design

With both an innovative and conservative version of each of the 50 sentences, there were 100 tokens for participants to listen to. These were ordered so that both the speakers and sentence types rotated in a regularised fashion and variables alternated from conservative to innovative. The first half of the experiment was repeated with inverse realisations for the second half, so that every participant heard each sentence twice, but with alternative variants. To minimise the effect that order might have on the experiment, two trials were conducted that were identical except that all conservative and innovative variants were inverted.

The experiment was conducted in a small quiet room on campus, over headphones on a laptop running Media Lab experimental software. All instructions were on the laptop, so that there was minimal interaction with the experimenter, who was always the author. Such considerations should help minimise any potential experimenter effects (Hay, Drager & Warren 2006). Participants were told that they were listening to actresses reading lines, and to focus on the voices of the women as opposed to what they actually said. This was to lessen the effect that the non-standard constructions might have on the age and social class ratings of the speakers.

It was stressed to participants that they needed to be as quick as possible and they were told that their response times were being taken. However, as participants were actually able to rate the age of a speaker before they had heard the entirety of a sentence, and because the /t/ was at the end of the sentence, this request was tempered by another, that they listen to each sentence in full before making their judgement.

Speakers would listen to each sentence while the age ratings were on the screen, with the title Please listen to the sentence in full then quickly estimate the age of the speaker. Age was given in five year blocks going from 15 to 40+, which made a six point scale, with the lowest number, 1, matched with the youngest age group, 15-19.

Once they had selected an age for the speaker, a new page would come up asking them to Please quickly estimate the social class of the same speaker. Social class was given on a four point scale, and next to the button '1' was the word 'higher', whilst next to the button '4' was the word 'lower'. This is different to the pilot study, where the titles 'working', 'lower-middle', 'upper-middle' and 'upper' were used, as it was decided that these labels were confusing/distressing for participants. They were not given any further explanation of what was meant by social class. They did not hear the sound file a second time when making the class ratings.

Participants were given a three sentence long trial before proceeding with the actual experiment. The sentences in the trial are given in below, and were from my archived recordings, discussed briefly in 3.2.1. The sentences, which would also be used in the practice run of the GJ task, ranged in grammaticality: Sentence (51) was ungrammatical, sentence (52) contained a non-standard construction (more earlier) and sentence (53) was grammatical.Three different males read the three sentences. They are aged 21, 24 and 48 respectively, and the first and last speakers were professionals.

Table 4.2: SP Experiment - Sex, age and class of participants in the Speaker Perception Study.

Trial	A	B
Total Participants	11	10
Total Females	7	3
Total Males	4	7
Min Age	18	18
Max Age	36	42
Median Age	28	22.5
Min Social Class Index	79	48
Max Social Class Index	150	15

51. Michael to America yesterday went.

52. It would be better if you could come over more earlier.

53. I havent had time to check my account yet.

After the practice examples, there was one more instruction screen, where participants had the opportunity to ask the experimenter if they had any questions or clarifications to make. Other than that, they would then proceed through the 100 sound files until they had finished.

4.1.5 Participants

Participants were recruited via signs around the University of Canterbury, a direct appeal to a first year linguistics class and by the occasional direct appeal to people known to the author. Twenty-one people participated in the experiment, and their class and age statistics are given in Table 4.2. The social class indices were calculated as described in Section 3.2.4.

4.2 Results

Responses were stored in Media Lab and accessed via Excel. With 21 participants listening to 100 sentences, there were 2100 total age and 2100 total class ratings. Before the raw data was analysed, instances where participants

had responded too early or too late were removed. There were 4 instances where the participant has responded too quickly (before the recording had finished), and a further 174 tokens where the participant had taken too long to respond. Reponse times were deemed overly long when they were two standard deviations over the mean response for a particular participant. If the removed token was an age rating, its alternate age rating and the class ratings were removed. If the removed token was a class rating, the alternate class rating was removed, but not the corresponding age ratings (because they had already been answered before the delay occurred). So a further 310 corresponding age and class ratings were removed, meaning that 488 tokens in total were excluded from analysis.

The remaining 3710 responses were analysed using linear regression models in R.2.0.0, using Harrells Design Library. Two linear regression models were fit by hand - one modelling perceived age, and one modelling perceived social class. The ratings for age and social class were considered in respect to information about the speakers, participants, the stimuli and the experiment. Which speaker had read the sentence was considered, in terms of individuals (SPEAKER), or alternately in terms of the age (SAGE) or NZSEI value (SCLASS) for that speaker, as were the age (AGEGROUP), social class (CLASS) and gender (GENDER) of the participant. From the stimuli we considered what type of sentence the speaker had read (TYPE), whether the realisation of this variable was the innovative or conservative version (REALISATION), the length of the track (TRACKLENGTH). Issues relating to the experiment were which trial group the participants were in (TRIAL), how far into the experiment the utterance was (ORDER), and how long the participant had taken to respond after hearing the track (RT). The factors that significantly affected the age ratings given to speakers are shown in Tables 4.3 and 4.4 (N=1935, R^2=0.306). Importantly for this study, the phrase final /t/ variant is significant, such that innovative forms (without release) receive lower age ratings than conservative ones (with release). Figure 4.1 shows how the individual speakers garnered significantly different age ratings, while Figure 4.2 shows how the different types of sentences affected the age ratings.

Table 4.3: SP Experiment - ANOVA for age ratings

Factor	d.f.	Partial SS	MS	F	P
REALISATION	1	15.46522	15.46522	18.88	<.0001
SPEAKER	4	526.6708	131.6677	160.78	<.0001
ORDER	1	31.31941	31.31941	38.24	<.0001
TRACKLENGTH	1	6.041196	6.041196	7.38	0.0067
RT	1	65.91032	65.91032	80.48	<.0001
TYPE	4	53.56609	13.39152	16.35	<.0001
regression	12 707.5547	58.96289	72	<.0001	
error	1922	1574.018	0.818948		

Table 4.4: SP Experiment - Coefficient table for age ratings

Coefficients	Value	Std.Error	t	p-value
Intercept	1.788995	0.165	10.81605	0.000
REALISATION=innovative	-0.18549	0.0427	-4.3456	<.0001
SPEAKER=E	0.001287	0.0663	0.01942	0.985
SPEAKER=K	1.480336	0.0659	22.47437	0.000
SPEAKER=L	0.586337	0.0647	9.0596	0.000
SPEAKER=S	0.380548	0.0645	5.89608	<.0001
ORDER	0.002243	0.000363	6.18413	<.0001
TRACKLENGTH	-0.00018	0.0000656	-2.71602	0.0067
REALRT	0.000137	0.0000153	8.97116	0.000
TYPE=COME	0.16927	0.0652	2.59823	0.0095
TYPE=DONE	0.10739	0.0676	1.58855	0.112
TYPE=GOT	0.382039	0.0687	5.56172	<.0001
TYPE=NORMAL	0.44224	0.0658	6.7258	<.0001

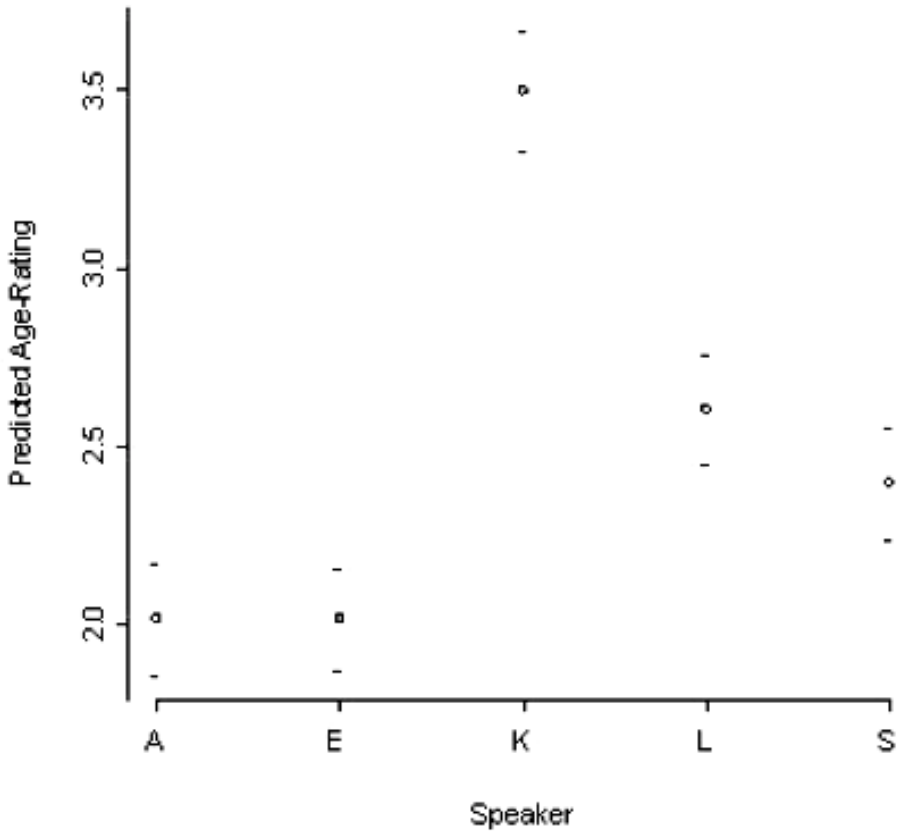

Figure 4.1: SP Experiment - Effect of SPEAKER on age rating

We see similar factors influencing the class ratings as shown in Tables 4.5 and 4.6 (N=1777, R^2=0.217). Again, REALISATION has an effect, such that the speakers are rated as belonging to a higher social class with the conservative variant, though it is a weaker factor than it was for the age ratings. Figure 4.3 shows the effect of speaker, which maintains a similar hierarchy to the age ratings, though speaker S, who had rated above speakers A and E for age, has the lowest class rating. The effect of the sentence type is shown in Figure 4.4, and is similar to Figure 4.2 modeling the age ratings, but here the BAD sentences receive higher class ratings than the preterite sentences.

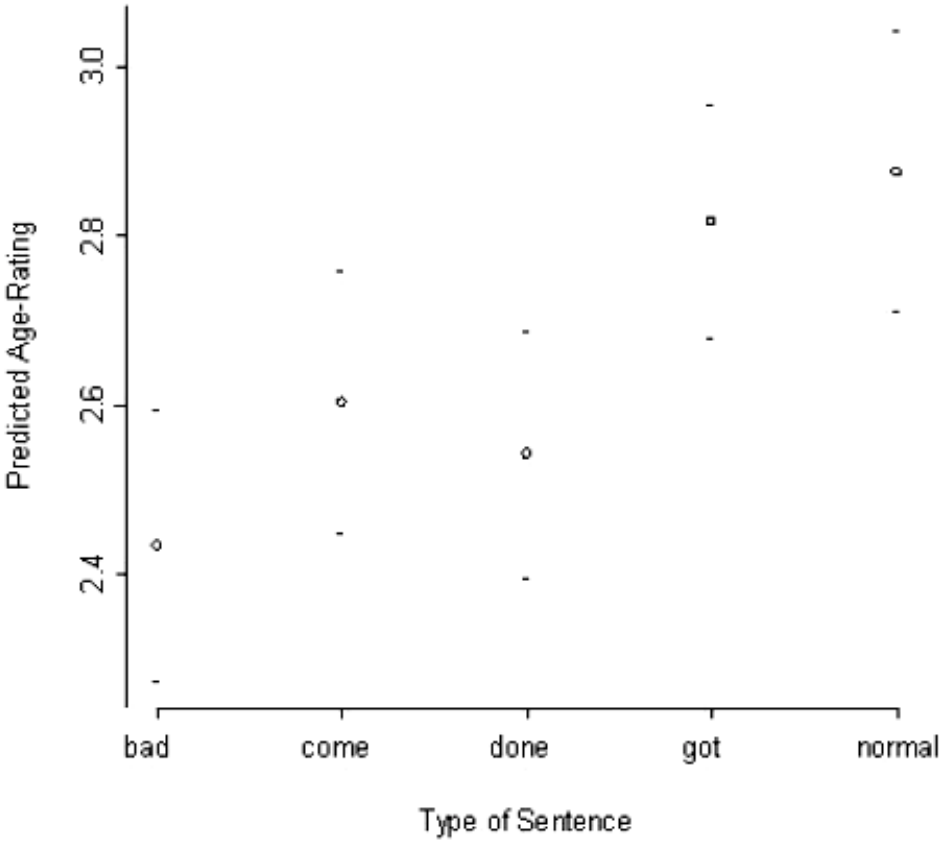

Figure 4.2: SP Experiment - Effect of TYPE on age rating

Table 4.5: SP Experiment - ANOVA for class ratings

Factor	d.f.	Partial SS	MS	F	P
SEX	1	2.527608	2.527608	4.77	0.0291
REALISATION	1	2.649327	2.649327	5	0.0254
SPEAKER	4	39.02695	9.756737	18.42	<.0001
TRIAL	1	6.668645	6.668645	12.59	0.0004
TYPE	4	208.2182	52.05454	98.28	<.0001
regression	11	263.3934	23.94485	45.21	<.0001
error	1764	934.2733	0.529633		

Table 4.6: SP Experiment - Coefficient table for class ratings

Coefficients	Value	Std.Error	t	p-value
Intercept	2.45089	0.07541	32.4989	0.000
SEX=m	-0.0803	0.03676	-2.1846	0.0291
REALISATION=innovative	-0.07725	0.03454	-2.2366	0.0254
SPEAKER=E	0.01308	0.05405	0.2421	0.809
SPEAKER=K	0.32953	0.05549	5.9386	3.45E-09
SPEAKER=L	0.15723	0.05324	2.9535	0.0032
SPEAKER=S	-0.11172	0.05451	-2.0495	0.0406
TRIAL	-0.13039	0.03675	-3.5484	0.0004
TYPE=COME	-0.04181	0.05468	-0.7645	0.445
TYPE=DONE	-0.25097	0.05535	-4.5344	<.0001
TYPE=GOT	0.39385	0.05525	7.1287	<.0001
TYPE=NORMAL	0.6999	0.05466	12.8039	0.0000

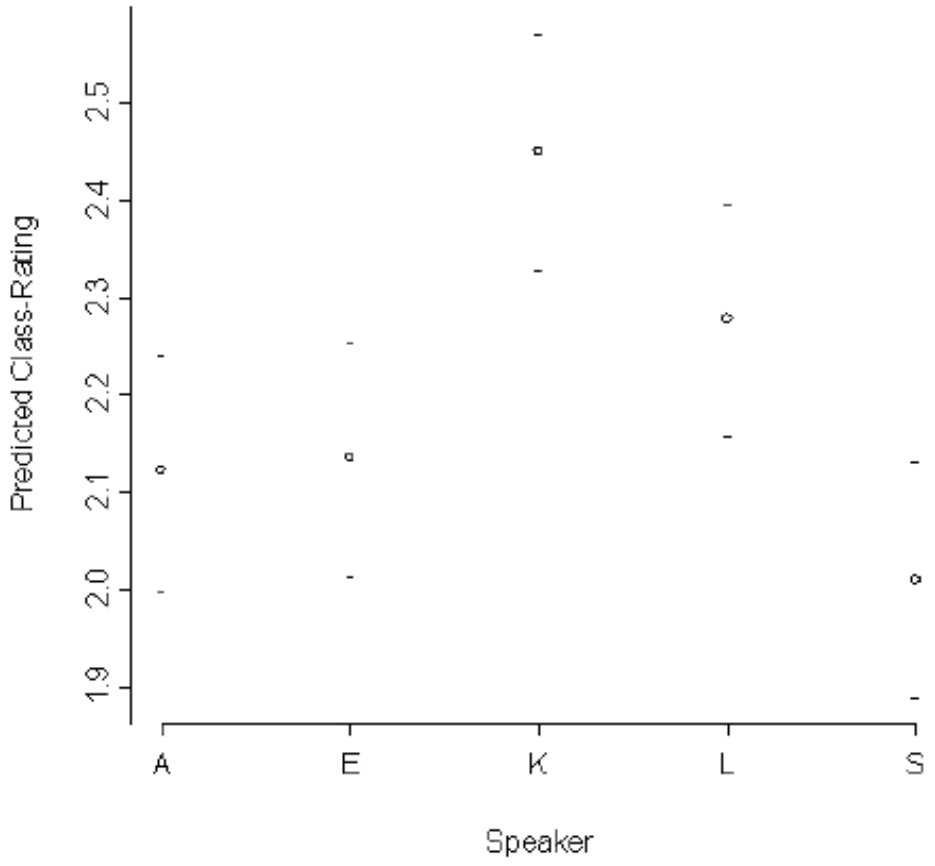

Figure 4.3: SP Experiment - Effect of SPEAKER on class rating

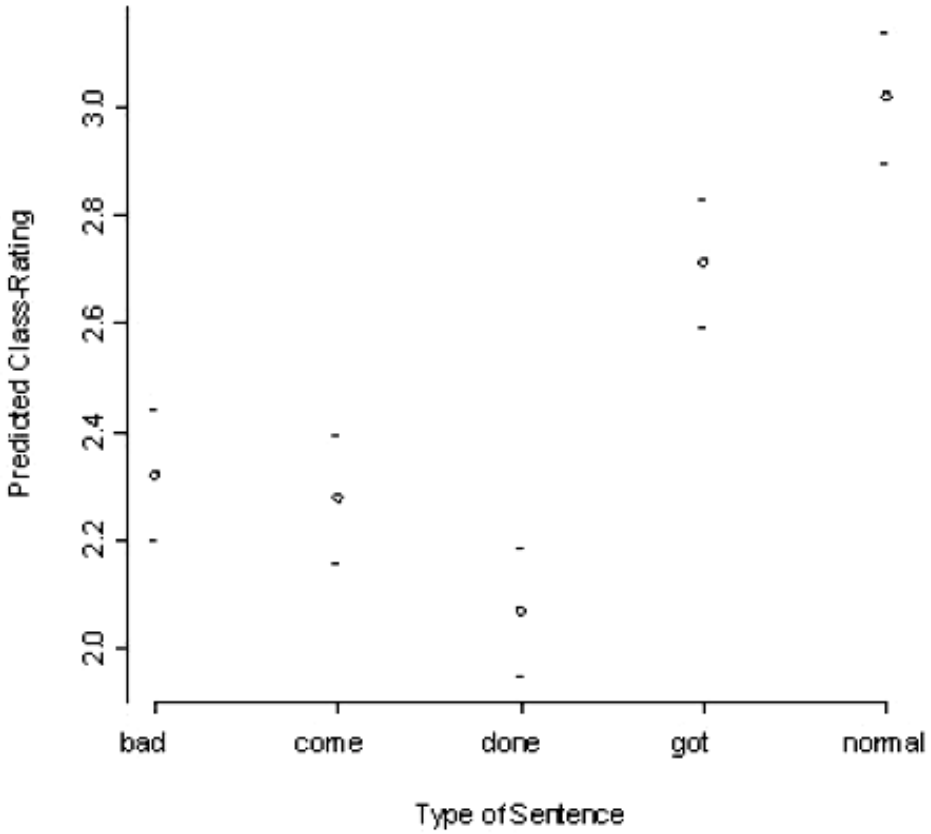

Figure 4.4: SP Experiment - Effect of TYPE on class rating

Table 4.7: Average age and class ratings given to the innovative and conservative realisations of all sentences, and the significant p-values of the differences from Wilcoxon matched pair tests.

	Age	Class
Innovative	2.547521	2.378378
Conservative	2.693899	2.454955
p-value (Wilcoxon)	<0.0001	0.003

Table 4.8: Breakdown of average age ratings (to 2dp) given to the innovative and conservative realisations by sentence type, and the p-value of the differences from Wilcoxon matched pair tests. Significant p-values are in bold

	BAD-age	COME-age	DONE-age	GOT-age	NORMAL-age
Innovative	2.34	2.69	2.69	2.44	2.72
Conservative	2.47	2.85	2.85	2.59	2.95
p-value (Wilcoxon)	0.0552	0.2565	**0.0261**	**0.0004**	**0.005**

The responses to the conservative and innovative variants were also run through Wilcoxon matched pair tests. The overall average responses and p-values are given below in Table 4.7. While both the age and class ratings given to innovative and conservative tokens were significantly different, the difference was more significant in the age ratings. Tables 4.8 and 4.9 show the averages and respective p-values of the age and class ratings given to the different types of sentences. While there was no interaction between type and realisation in the logistic regression models presented earlier, we can see that the effect was stronger for certain types of sentences than others. There was a significant difference between the age ratings of the conservative and innovative tokens given to all the types of sentences except the COME sentences (the BAD sentences are borderline significant). For the class ratings, the differences between the innovative and conservative ratings were only significant for the GOT and DONE sentences.

Table 4.9: Breakdown of average class ratings (to 2dp) given to the innovative and conservative realisations by sentence type, and the p-value of the differences from Wilcoxon matched pair tests. Significant p-values are in bold

	BAD-class	COME-class	DONE-class	GOT-class	NORMAL-class
Innovative	2.44	2.92	2.18	1.94	2.59
Conservative	2.24	2.98	2.25	2.07	2.71
p-value (Wilcoxon)	0.9912	0.3642	**0.03823**	**0.02332**	0.2601

4.2.1 Summary of Results

Whether a sentence had contained the innovative or conservative phrase final /t/ variant had a significant effect on both the age and social class ratings given to the speakers in this experiment. The effect worked so that when participants heard a sentence with an innovative variant, they would rate the speaker as younger and of a lower social class than when they heard the conservative variant. There was also similar effect for the sentence type, such that speakers were rated as older and of a higher social class when they were reading the NORMAL and GOT sentences, compared to when they read the BAD, COME and DONE sentences. While the BAD sentences generally garnered the youngest age ratings, the preterite constructions garnered the lowest class ratings. This is all despite the fact that participants had been told repeatedly to only base their judgments on the voice of the speaker, and not what they said.

Another factor that was the significant in both the age and class ratings was which of the five speakers was reading the passage. Speaker K received the highest age and class ratings, followed by speaker L. The two women were in fact the oldest of the speakers, though they both also had significantly lower NZSEI scores than the others. This could reflect problems with how NZSEI scores are calculated, the fact that the women have had theatrical training, but also the fact that listeners associate the speech of older speakers with speech from speakers of a higher social class. However, speaker S received higher age ratings than speakers A and E, but not higher class ratings. The age ratings were affected by the order and tracklength of a token, and the length of time it took participants to respond. Sentences that participants encountered further into the experiment received higher age ratings than ear-

lier ones, which is possibly due to pressure participants felt to use more of the age continuum even though the speakers were all relatively young. The longer a token was, the younger the speaker would be rated, which could be due to an increase in accuracy the more participants had to base their judgments on. The longer participants took to respond, the higher they would rate the age of the speaker. This could mean that when in doubt, participants erred on guessing higher.

Additional factors in the social class ratings were the sex of the participant and the trial that they had been in. Males rated the speakers as belonging to a lower social class than did females, and participants in Trial 1 rated the speakers as belonging to a higher social class than did those in Trial 2. This could well be due to the fact that there were considerably more females in the first trial compared to the second, though the model should have held gender constant.

The results from the Wilcoxon matched pair tests, which compared the responses to the conservative realisations with the responses to the innovative sentences, not only overall, but within each sentence type, showed that the age ratings had been more greatly manipulated than the class ratings. While it is tempting to assume that this means the phrase final /t/ is a more salient age than class marker, there were also complicating experimental factors: participants not only rated class after they rated age (and thus longer after they had heard the recording), but the 4 point class continuum might have resulted in less subtle variability that the 6 point age continuum. Furthermore, the rating of class has negative social connotations the way that the rating of age does not, and participants may have hesitated to behave critically.

4.3 Discussion

While the documented distribution of phrase final /t/ would predict the direction of the observed effect, they would not necessarily predict that the /t/ would be able to change the age and class ratings of a speaker in sentences full of other socially salient features of NZE. The middling results of the pilot suggest that not all socially variable features would be so effective, and it

is the binary (released or unreleased) distinction, marked turbulence of the release and phrase final position that may make the /t/ particularly salient.

The fact that the direction of the SP Experiment results mirrors the production of the variants adds to mounting evidence that encountered realizations of phonemes are stored complete with speaker information, as embodied in Exemplar Theory (Johnson 1997, Pierrehumbert 2006)). If listeners will rate someone as being older because they hear a released /t/, this must be because they are accessing memories of other times they have heard a released /t/, and those memories must include attributes of the speaker, such as their age and class. It would not be the case that participants had never heard a younger, non-professional speaker release phrase final /t/, but rather that more of the stored memories were tagged with older and professional attributes, and it is the higher probability that leads to the strong association.

As mentioned earlier, it is probably unwise to interpret the results as conclusive evidence that the presence or absence of a /t/ release is more meaningful in terms of the age of the speaker as opposed to their social class, since this was not what the experiment was testing. The pilot suggests that /t/ may in fact be a stronger indicator of class.

It is also open for interpretation whether both realisations of the variable were having equal and opposing effects (such that one made people think the speaker was younger and the other that they were older), or it could be, as (Campbell-Kibler 2007) has explored, that one of the variants is meaningful in a way that the other isnt. For example, the unreleased /t/, being considerably more common than the released /t/, could actually be fairly neutral in terms of social associations, and the difference is coming entirely from the strong associations of the released variant with older, professional speakers. Similarly, it is possible that the released /t/ has strong, say, professional connotations, whereas the unreleased variant has strong associations with younger speakers, so that both variants carry social meaning, but not necessarily converse meanings.

The most puzzling part of the data is the fact that the realisation effect is not as significant in the BAD and especially the COME sentences. Although there was no interaction of TYPE with REALISATION in the overall linear regression model, the differences in ratings given to the conservative and

innovative variants did not come out as significant for these sentences in Wilcoxon matched pair tests (though there was an almost significant effect in the age ratings for the BAD sentences). It could be the case that the non-standard constructions in the sentence are overriding any effect that the /t/ might have, but then we would also expect to see a null effect in the DONE sentences, and in fact the difference is significant in both the age and class ratings of the DONE sentences.

Finally, the effect of the construction type is relevant to the primary question of my thesis. Much like the effect of the variable, we saw that when speakers used constructions that are produced mostly by younger non-professionals, the speakers were rated as younger and of a lower social class than when they used more standard forms. Such findings could have two interpretations: the first, that constructions are stored with speaker information, the second, that participants, regardless of their experience, associate less grammatical sentences with younger and non-professional speakers.

In this argument, the behaviour of the BAD sentences is pivotal. The BAD sentences are non-grammatical, that is, they are not something we would expect any native speaker of NZE to have some across. They should have no social meaning because they are not produced by any speaker groups. What we see is that the BAD sentences pattern closely with the non-standard preterites. In the age ratings, the BAD sentences receive the lowest age ratings overall, whereas in the class ratings, they are close to but above the preterite sentences. This suggests that both explanations are to some extent correct. People will rate speakers younger in the BAD sentences because of a prejudice that says that young people use more ungrammatical forms than older speakers. However, in rating the class of the speakers, they will give the lowest ratings to the sentences with preterite COME and DONE, because, beyond the prejudice, they have encountered non-professionals using the forms. This is an interesting pattern that conceivably could be replicated in the GJ task.

4.4 Conclusion

The results of this study show that listeners can and do use specific phonetic detail in judging aspects of a speakers social identity. Furthermore,

this detail can be embedded in a sentence filled with various other features and still have an effect. This is particularly impressive when the sentences already contain constructions that carry strong social information, and indicates that subtly altering phonetic detail may also be able to alter the perception of such constructions, as we test in the following chapter.

Chapter 5

Grammaticality Judgement Experiment

5.1 Introduction

The primary aim of this thesis is to examine the effect that socially meaningful phonetic detail can have on the grammaticality ratings given to sentences that contain socially meaningful morpho-syntactic forms. The experiment described in the previous chapter showed that phrase final /t/ appears to be a salient and reliable phonetic variable in speaker perception, and its manipulation can alter both the perceived age and social class of a speaker. The experiment described in this chapter tests whether, because of these social associations, the realisation of phrase final /t/ can also alter the perceived grammaticality of sentences.

Many details of the methodology mirror that of the SP Experiment, and the same sentences are used in both (though there are also additional filler sentences here). However, unlike the SP Experiment, where participants were asked to rate speaker attributes, here the focus was on the constructions themselves.

There are a number of possible outcomes of such an experiment. The null hypothesis is that there is no difference in the grammaticality ratings given to sentences with the conservative and the innovative realisation of the /t/. An experience-based hypothesis would predict that there would be a difference in

the ratings for the two realisations in the target sentences only, such that being produced more by younger, more innovative speakers, they will be rated as better when presented with the innovative variant. There should be no effect of realisation on the control sentences. A prejudice-based hypothesis could have two outcomes. Firstly, participants might associate conservative speakers with more correct and grammatical speech, and thus rate all the sentences with the released /t/ as more grammatical than those with the unreleased /t/. Alternatively, participants might associate conservative speakers with correct and grammatical speech, but innovative speakers with innovative and incorrect speech, so that they rate all less grammatical sentences, even the BAD filler sentences, as better with the innovative variant, whereas they rate the NORMAL filler sentences best with the conservative variant.

This chapter proceeds as follows. In 5.2, the filler sentences are introduced, then in 5.3, the experiment design is outlined. The results of the experiment are described in 5.5, and discussed in 5.6.

5.2 Sentences

The 50 pairs of target and control sentences used in the SP Experiment were also used in this experiment. To take the attention away from these constructions and make the grammaticality rating task more diverse, 60 unmanipulated filler sentences were created. They consist of NORMAL sentences, BAD sentences and BORDERLINE sentences.

5.2.1 Normal Sentences

As with the normal control sentences introduced in 4.1.1, the 20 normal fillers contained standard constructions wiht no known social biases. They are shown below (1-20). It was predicted that these sentences, along with the normal controls, would receive the highest grammaticality ratings.

1. David cooked dinner and I did the dishes.
2. I don't know when she'll be home.
3. It's so nice to have the weekend off.

4. He said that he didnt mind either way.
5. I'm looking forward to a proper holiday.
6. He's actually very intelligent.
7. Mary's been trying to get me to eat better.
8. It's the best university in the country.
9. We have a lot of friends and family here.
10. She's under a lot of pressure at the moment.
11. Katy asked me to go to dancing lessons with her.
12. He's really respected by everyone at work.
13. We went for a walk on the beach after our swim.
14. I'm just thankful it wasn't any worse.
15. Frank handed in his resignation today.
16. She wasn't as bad as the last teacher they had.
17. Jess was laughing so hard she was crying.
18. This is my favourite song at the moment.
19. The computer kept crashing and it was driving me mad.
20. I've been waiting for over an hour.

5.2.2 Bad Sentences

The bad control sentences described in 4.1.1 contained constructions that contained non-native like mistakes, while still maintaining coherence. The 20 bad filler sentences introduced here contained the same sort of errors as the controls: pronouns with incorrect case (21-28), plural subjects with verbs marked for singular subjects (29-34) and verb finality (35-40). It was expected that these sentences would receive the

lowest ratings, and that their presence would force ratings of the non-standard, but experienced, constructions away from the ungrammatical periphery because these sentences were so much worse.

21. When us asked she, her admitted everything.
22. Him's been busy at this computer all morning.
23. I'm committed to giving they all the help them need.
24. For Christmas, them bought she a new stereo.
25. Them said us would have to wait.
26. Him obviously didn't think of she.
27. Her decided to leave it up to we.
28. When they heard the news, they was angry and upset.
29. They thinks dinner will be too late.
30. We hopes Jimmy doesnt end up in court.
31. They argues almost every night.
32. We really loves going out to dinner.
33. They is looking at buying a house in the city.
34. We is considering all our options.
35. John to the shops after work went.
36. I to sing at the concert hope.
37. After lunch, I to the bank am going.
38. The garden much better now is looking.
39. Henry to become a policeman is training.
40. My family a traditional Xmas dinner always have.

5.2.3 Borderline Sentences

Another twenty sentences were included in the experiment that consisted of another two types of construction that show social variation in NZE. This was so there were more sentences other than the targets which we would expect to receive ratings somewhere in between the polarities of grammatical and ungrammatical. The first ten sentences all contained non-standard coordinated pronoun constructions (41-50). Prescriptively, me should appear in subject position and me as the object of a verb or preposition. However, with conjoined pronouns, the form of the pronoun appears to be more influenced by its position in the coordinate, with me favouring the initial position and I better when second (Quinn 2005). Angermeyer and Singler (2003) suggested that there were two forms, the vernacular (*me and Tim*) and the polite (*Tim and I*), and they found that older speakers preferred the latter to the former, whilst younger speakers were the opposite. Five of the co-ordinated pronoun sentences contained *me and X* subjects, and in the other five *X and I* followed the preposition *to*.

41. Me and Laura haven't talked in weeks.
42. Me and Tim have been shopping all day for a present.
43. Me and mum get along pretty well most of the time.
44. Me and George went there last week for dinner.
45. He offered the place to Greg and I but we said no.
46. Dad gave the car to Sarah and I to share.
47. She promised the job to Sam and I but then gave it to them.
48. Yesterday Hannah talked to Henry and I about leaving.
49. Me and John have been cycling every weekend.
50. My grandmother gave a necklace each to my cousin and I.

The next ten sentences contained non-standard negated modal+HAVE constructions. Typically, negation of such constructions occurs between

the modal and the HAVE, either in the full form of not or with nt clitising to the modal. However, Walker (2005a) reports cases where the not follows a reduced form of have. In an aural task where participants were asked to rate whether the constructions were something that they themselves would use, that some New Zealanders would use, but they wouldnt, or whether no native speaker of NZE would say them, the non-standard variant was more readily accepted with the epistemic modals may, must and might than with the deontic modals should and would. The results of Walker (2005a) confirmed that may was most preferred and should least so, but the ratings of must and would interacted with the social class of the participant, so that professionals rated must of not as more grammatical than would of not, while the non-professionals had it the other way round.

The ten sentences created for this experiment contained two sentences of each of the modals would, must, may, might and should. Each modal appeared with a first and a third person singular pronoun. The sentences are shown below (51-60).

51. I would've not danced with Jacob.
52. She would've not helped even if youd asked.
53. I must've not understood him.
54. He must've not been listening.
55. I may've turned the oven off.
56. She may've not received the letter.
57. I might've not been in the room at the time.
58. She might've not checked her inbox today.
59. She should've not been spying on us.
60. I should've not trusted Jenny.

5.3 Speakers

The same five young women who recorded the sentences for the SP Experiment (see 4.1.2) returned to record the new sentences for this experiment. They were again recorded individually into Sony Sound Forge via a head mounted microphone connected to a USB pre. Two unique sentences were selected from each speakers recordings for each of the BORDERLINE constructions, whilst a further four were taken from both the remaining GOOD and BAD sentences.

5.4 Experiment Design

There were 110 unique sentences in total for the experiment, but the 50 with phrase final /t/ repeated so that both the conservative and innovative versions were heard, making a total of 160 sentences for participants to rate. The sentence types rotated in a regular fashion. Care was taken to have the done and come constructions apart from each other, and to ensure that target constructions did not always follow the same type of construction, i.e., good or bad sentences. There was also a rotation of the type of realisation, which went innovative then conservative, but was occasionally overridden because alternating the realisations for each of the sentence types (so that a COME sentence always had an innovative then conservative realisation) was considered more important. Finally, the speakers also moved through their own, independent rotation.

One version of each of the fifty unique /t/ sentences was heard within this half of the experiment. In the second half, the experiment repeated itself, except that realisations of the /t/ in the relevant sentences were inverted (so that what was conservative was now innovative, etc), and that the filler sentences were all new and unique (but of the same type as in the first half of the experiment).

To counteract the effect that order might have on the responses, the participants were split between one of two groups which were identical except that in the second trial the realisations were converse to those of the first trial.

Participants were asked to rate the grammaticality of the sentences on a

six point scale, where 6 meant that they thought a sentence was completely grammatical and 1 meant they thought it was completely ungrammatical. No specific explanation of what the term grammatical meant was given, though in the instructions it was emphasised that our interest was in native speaker intuitions about language, and not grammar as it is taught in schools.

The experiment was again run off Media Lab, off a laptop in a small, quiet room on campus. After manually filling out an information sheet about themselves, participants were put on the computer and given headphones. All instructions appeared to them written on the screen, spread over three screens. As well as stressing that they should rate the sentences on how they felt, they were also asked to listen to the sentences in full before responding, and bearing that in mind, to be as quick as they could, as their response times were being recorded. They then did a small trial run, which consisted of the same three sentences used in the SP Experiment (4.1.4).

After they had done the trial, there was one final instruction page telling them that if they needed to make any clarifications, this was the time to ask the experimenter. Other than that, after reiterating that they needed to be quick but at the same time to listen to the whole sentence before responding, they were free to begin the experiment. Responses and reaction times were automatically collected by the program.

5.5 Participants

Participants were recruited by signs around the University of Canterbury and a direct appeal to first year linguistics and engineering classes asking for volunteers. They were given $5 and a chocolate fish for their participation in the experiment, which generally took less than half an hour. The breakdown of the participants is given in Table 5.1. Note that there are considerably less males in Trial A, and the Social Class Indices are a little lower in Trial B.

Table 5.1: GJ Experiment - Sex, age and class of participants

Trial	A	B
Total Participants	14	16
Total Females	11	8
Total Males	3	8
Min Age	18	18
Max Age	28	24
Median Age	19.5	19
Min Social Class Index	60	50
Max Social Class Index	154	148
Median Social Class Index	102.5	88

5.6 Results

The 4800 responses from all 30 participants were exported from Media Lab into Excel. From here, 34 tokens were removed because participants had responded before the sound file was through, and for the 19 of those that had been manipulated, their alternate versions were also removed. A further 185 tokens were removed because participants had taken too long to respond. Responses were deemed overly long when they were more than two standard deviations over the mean response time for each participant, for each sentence type (due to the fact that response time was largely influenced by which type of sentence participants were listening to). In the 118 instances where these tokens had been manipulated, their alternate partner was also removed from analysis. In total then, 592 tokens were exluded, leaving 4444 responses for analysis.

Before statistical analysis of the data, it is worth looking at the basic patterns of the sentences. Table 5.2 shows the average grammaticality ratings and response times given to each of the different types of sentences. Unsurprisngly, the BAD sentences were rated the least grammatical, and the NORMAL sentences were rated the most grammatical. The GOT sentences were also rated highly, followed by the BORDERLINE sentences. COME sentences were rated a little higher than DONE sentences. Participants were fastest in responding to the BAD sentences, then the DONE and NORMAL sentences, then the COME, GOT and BORDERLINE sentences.

Table 5.2: GJ Experiment - Average grammaticality ratings (to 2dp) and response times broken down by sentence type.

	BAD	DONE	COME	BORDERLINE	GOT	NORMAL
Rating	2.04	2.95	3.4	4.47	5.07	5.60
RT (ms)	1069.608	1312.381	1517.406	1587.831	1572.023	1314.355

5.6.1 Overall Factors affecting Grammaticality Ratings

A linear regression model was hand fit and run in R.2.0.0, using Harrells Design Library. The grammaticality ratings were the dependent variable. As we are primarily interested in the effect of the phrase final /t/, this model only includes data from the target and control sentences (2726 tokens). Independent variables tested concerned the participant, the speaker, the sentences or the experiment. The participants age (AGE), NZSEI score (CLASS) and sex (SEX) were in the model, as well as the trial they were in (TRIAL). The individual speaker (SPEAKER) was included in the model, though this could be substituted with either the average age (SAVERAGEAGE) or class (SAVERAGECLASS) rating they received from the Speaker Perception Experiment so as to better explain the differences between the speakers. The average age rating (AGERATING) and class rating (CLASSRATING) given to each sentence in Experiment 1 were also alternatively tried in the model. Other included factors were the type of sentence participants heard (TYPE), whether the sentence was a target sentence or not (TARGET), the realisation of the phrase final /t/ in manipulated sentences (REALISATION), the length of the recording (TRACKLENGTH), and the placement of the token within the experiment (ORDER).

Tables 5.3 and 5.4 (N=2726, R^2=0.58) and show the factors that were found to be significant in the overall grammaticality ratings given to the manipulated sentences. Unsurprisingly, the type of sentences being rated was highly significant. NORMAL and GOT sentences were rated best, and then the COME, DONE and lastly the BAD sentences. Speaker also had an effect, but in an interaction with the type of sentence (Figure 5.1). For example, Speaker K received the highest grammaticality ratings in the NORMAL and GOT sentences (and second highest with the BAD sentences), but she re-

Table 5.3: GJ Experiment - ANOVA model for overall factors on grammaticality ratings

Factor	d.f.	Partial SS	MS	F	P
AGERATING	1	7.73	7.73	5.75	0.02
TYPE	20	4433.35	221.67	164.87	0.00
All Interactions	16	178.53	11.16	8.30	0.00
SPEAKER	20	255.98	12.80	9.52	0.00
All Interactions	16	178.53	11.16	8.30	0.00
SEX	1	13.64	13.64	10.15	0.00
AGE	1	32.80	32.80	24.40	0.00
TRIAL	1	12.30	12.30	9.15	0.00
TRACKLENGTH	1	12.43	12.43	9.24	0.00
TYPE * SPEAKER	16	178.53	11.16	8.30	0.00
Regression	29	4909.71	169.30	125.92	0.00
Error	2696	3624.68	1.34		

ceived the lowest ratings for the COME sentences. Speaker E, comparatively, rated well with preterite sentences, but comparatively poorly with the BAD, NORMAL and GOT sentences. This seems to reflect the perceived age and class of the speakers (refer to Figures 4.1 and 4.3).

This model also includes the average age rating each token received in the SP Experiment, and the negative coefficient value tells us that, for each sentence, the older the speaker had been rated on average in the SP Experiment, the less grammatical that sentence was rated. Speaker and sentence type were factors on the age ratings given to tokens in the SP Experiment, but as the AGERATING effect is significant in a model that already takes both of these factors into account, we must assume that AGERATING is capturing sentence specific differences. Amongst these could be the realisation of the variable, which also significantly affected age ratings, though REALISATION is not a significant factor in this model. It should also be noted that we can substitute CLASSRATING for AGERATING for an inferior model. Both cannot be tested in the same statistical model due to the high degree of collinearity between them.

Table 5.4: GJ Experiment - Coefficient Table for grammaticality ratings

Coefficients	Value	Std.Error	t	p-value
Intercept	2.56853	0.26832	9.5725	0.00
SPEAKER=E	0.54667	0.16006	3.4153	0.0006
SPEAKER=K	1.34264	0.1772	7.5768	<.0001
SPEAKER=L	0.9106	0.16493	5.5213	<.0001
SPEAKER=S	1.61445	0.16852	9.5799	0.00
TYPE=COME	1.8731	0.162	11.5627	0.00
TYPE=DONE	1.27611	0.15965	7.9932	<.0001
TYPE=GOT	3.89021	0.1606	24.2234	0.00
TYPE=NORMAL	4.32878	0.16311	26.5397	0.00
AGERATING	-0.15175	0.06991	-2.1707	0.03
SEX=m	-0.16344	0.05145	-3.1766	0.0015
TRIAL=B	0.15356	0.0507	3.0286	0.0025
AGE	-0.0478	0.00967	-4.9439	<.0001
SPEAKER=E*TYPE=COME	0.08092	0.22445	0.3605	0.719
SPEAKER=K*TYPE=COME	-1.29437	0.22767	-5.6854	<.0001
SPEAKER=L*TYPE=COME	-0.52439	0.22489	-2.3317	0.02
SPEAKER=S*TYPE=COME	-1.19755	0.23143	-5.1745	<.0001
SPEAKER=E*TYPE=DONE	0.09642	0.22346	0.4315	0.666
SPEAKER=K*TYPE=DONE	-0.82769	0.22475	-3.6827	0.0002
SPEAKER=L*TYPE=DONE	-0.33142	0.22465	-1.4752	0.14
SPEAKER=S*TYPE=DONE	-1.17507	0.22724	-5.1711	<.0001
SPEAKER=E*TYPE=GOT	-0.96252	0.22452	-4.287	<.0001
SPEAKER=K*TYPE=GOT	-1.27789	0.23507	-5.4363	<.0001
SPEAKER=L*TYPE=GOT	-0.79184	0.22404	-3.5343	0.0004
SPEAKER=S*TYPE=GOT	-1.49959	0.22604	-6.634	<.0001
SPEAKER=E*TYPE=NORMAL	-0.71899	0.22716	-3.1651	0.0016
SPEAKER=K*TYPE=NORMAL	-1.05276	0.22421	-4.6953	<.0001
SPEAKER=L*TYPE=NORMAL	-0.94135	0.23104	-4.0743	<.0001
SPEAKER=S*TYPE=NORMAL	-1.49953	0.22729	-6.5974	<.0001

Figure 5.1: GJ Experiment - Interaction of TYPE and SPEAKER on grammaticality ratings

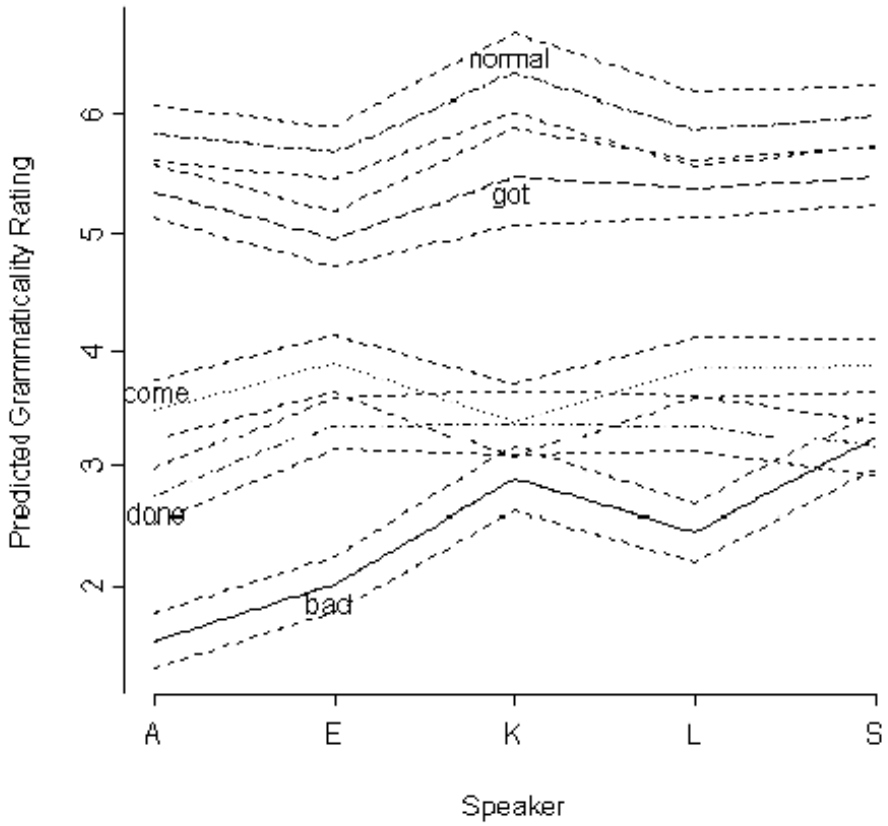

Table 5.5: Average grammaticality ratings (to 2dp) given to the conservative and innovative realisations of manipulated sentences.

	BAD	DONE	COME	GOT	NORMAL
Conservative	2.11	3.37	2.98	5	5.59
Innovative	2.14	3.42	2.92	5.14	5.56

There were also a number of participant effects that came out as significant. Male and older speakers rated sentences as less grammatical than female and younger participants. Participants in Trial B rated sentences as better than those in Trial A.

A further series of models looking at the influential factors on each of the sentence types separately showed that AGERATING was only significant for the come sentences (p-value=<0.001), in a model that included SPEAKER, AGE and TRACKLENGTH). This implies that the effect, where younger sounding tokens are rated better, is strongest in the COME sentences, which younger speakers use more than older speakers.

As with the SP experiment, a series of Wilcoxon matched pair tests were run as another way to test whether the differences between ratings given to the conservative and innovative versions were significant. They were not, for any of the types of sentences, though for the GOT sentences the difference was almost significant (p=0.07), and went such that the innovative was rated better that the conservative. The average ratings given to each sentence type for the conservative and innovative versions are in Table 5.5.

5.6.2 RATEDIFF

To explicitly explore the differences, if any, in the responses given to the conservative and innovative realisations of the same sentence, another linear regression model was run which looked only at the manipulated sentences, and had as the dependent variable the difference in the ratings given to the conservative and innovative version of the same sentence (RATEDIFF). That is to say, for each sentence, the rating each participant gave the innovative version was subtracted from the rating the same participant gave the conservative version of the same sentence. A positive value for RATEDIFF would

mean that the conservative version had been rated as more grammatical, a negative rating that the innovative version had been rated as more grammatical, and a value of 0 would mean that there was no difference between the two versions. For anything to come out as significant in these models, there must be a principled difference in the ratings given to the different variants.

There were 2726 tokens included in the linear regression model described in the previous section. As there can only be one RATEDIFF score for every pair, the number of tokens including in the RATEDIFF model halved to 1363. Many of the same independent variables included in the other model were tested in this one as well: participant factors (SEX, AGE, CLASS, TRIAL), speaker factors (SPEAKER, SAVERAGEAGE, SAVERAGECLASS), experiment factors (ORDER) and token factors (TYPE, TARGET, TRACKLENGTH). Several new factors were also included: which realisation participants had heard first (FIRSTWAS); the length of the release in the conservative realisation (LENGTHT); the difference in the length of time it took to respond to each version (TIMEDIFF).

Two other related factors were alternately tried in the models aswell: CLASSRATEDIFF and AGERATEDIFF. Each sentence had a CLASSRATEDIFF and an AGERATEDIFF value, which were calculated by subtracting the averageage or class rating given to the innovative version of a sentence from the age or class rating given to the conservative version of the same sentence in SP Experiment. Like RATEDIFF, a positive value for these new variables meant that the speaker had been rated as older or of a higher social class with the conservative variant, a negative value meant that they had been rated older or of a higher social class with the innovative variant, and a 0 value meant that there had been no difference between the ratings for the variants. The inclusion of CLASSRATEDIFF and AGERATEDIFF in these models allows us to explicitly compare the impact that the phrase final /t/ manipulation had on the age and class ratings with the impact it had on the grammaticality ratings given to the same recording.

The factors that were influential in the size and direction of the RATEDIFF were which of the two variants participants had heard first, and the average CLASSRATEDIFF of the sentence in the S.P. Experiment (Tables 5.6 and 5.7 (N=1363, R^2=0.009)). Whichever variant participants heard a sentence with first, they rated that sentence better with the other variant.

Table 5.6: GJ Experiment - ANOVA for RATEDIFF

Factor	d.f.	Partial SS	MS	F	P
CLASSRATEDIFF	1	6.424696	6.424696	4.72	0.03
FIRSTWAS	1	13.37552	13.37552	9.82	0.0018
regression	2	19.59368	9.79684	7.2	0.0008
error	1360	1851.558	1.36144		

Table 5.7: GJ Experiment - Coefficient table for RATEDIFF

Coefficients	Value	Std.Error	t	p-value
Intercept	-0.09983	0.04595	-2.172	0.0299
CLASSRATEDIFF	-0.32011	0.14736	-2.172	0.0300
FIRSTWAS=innovative	0.19814	0.06321	3.134	0.0018

This seems to translate as an effect of order, such that the further into the experiment a token occurred, the better it was rated. However, there was no significant effect of order in the overall rating model described in 5.6.1. The CLASSRATEDIFF effect says that the more the speaker of the sentence was rated as being from a higher social class with the conservative relative to the innovative variant, the less grammatical the sentence was rated with the conservative relative to the innovative variant.

5.6.3 Summary of results

Unsurprisingly, the type of construction in the sentence was a major factor in the grammaticality ratings given to the sentence. NORMAL sentences were rated as the most grammatical, followed closely by the GOT sentences, then the BORDERLINE filler sentences. BAD sentences were rated as the least grammatical, then the DONE and then the COME sentences. In terms of response times, participants were fastest in making judgements on the BAD sentences, then the NORMAL and DONE sentences.

Who the speaker of a sentence was also affected grammaticality ratings and response times, in an interaction with the type of sentence that was being rated. Speaker K, rated as oldest and most professional sounding in the SP Experiment, was rated best with the NORMAL, GOT and BAD sentences

but worst with the COME sentences. On the other hand, Speakers E and L, who had relatively lower age and class ratings, were rated best with the COME sentences, but were rated worst with the NORMAL sentences. Thus the speaker differences appear to reflect the perceived age and class of the speakers, and the likelihood that such speakers would use the construction. However, it should be noted that the speakers all read different sentences, and so, while I do not believe this is the case, it is possible that, for example, Speaker K happened to be matched with the two least grammatical come constructions and Speaker E with the two most grammatical. Time constraints mean that I cannot explore this further in this thesis, though I intend to get the grammaticality ratings of the sentences when presented visually only, in order to check the validity of this argument.

There was an additional effect of the agerating the particular sentence received in the SP Experiment, such that the older the speaker was rated for a particular sentence, the worse the grammaticality rating of the same sentence. A closer look at the data suggested this effect was strongest in the COME sentences. As speaker and type were already included in the model, agerating was presumably getting at other inter-sentence variation, possibly including things like the realisation of the phrase final /t/. Thus, while REALISATION was not significant in the model, it is possible that it is being captured somewhat in AGERATING, though only when the realisation had succeeded in altering the perceived age of a speaker.

Participant factors were also important, and age and sex both had roles: older and male participants rated the sentences worse than younger females. Younger females generally lead language change (Labov 1990), so it could be that for, say the GOT and COME sentences, females rated these better and they are carrying the effect. Participants in the second trial rated sentences higher than those in the first trial.

As mentioned, the realisation of the phrase final /t/ did not have a significant effect in the overall model, and in Wilcoxon matched pair tests there was no significant difference between the ratings given to the conservative and innovative versions, though the effect was approaching significance for the GOT sentences. The GOT sentences are not only where we saw the strongest effect of realisation in the age and perhaps more importantly the class ratings in SP Experiment, but are also the only instance where the phrase final /t/ was

being manipulated on the actual construction in question. However, GOT sentences were also rated as highly grammatical, and there was less variation in responses, suggesting that responses may have frequently plateaued, leaving little room for movement.

When we focussed explicitly on the differences in ratings between the conservative and innovative versions of sentences by putting RATEDIFF into the model as the dependent variable, we found a significant effect of the CLASS-RATEDIFF, such that the more a speaker of a sentence had been rated as having a higher social class with the released /t/ relative to the unreleased /t/, the worse the grammaticality rating of the sentence with the released /t/. This is congruent with the AGERATING pattern seen in the overall model, where the older a speaker had been rated for a particular sentence, the worse that particular sentence was rated.

5.7 Discussion

The COME, DONE and HAVE-got sentences were included in the experiment as constructions that showed some socially related variation in NZE, based on the literature. While they were all rated between the BAD and NORMAL sentences in this experiment, the DONE sentences not only received low ratings nearing the BAD sentences, but participants responded as quickly to these sentences as they did to the NORMAL sentences. This may suggest that the preterite DONE constructions are becoming less and less frequent, and this in fact may be the case (Heidi Quinn, personal communication). At the other end, the HAVE-got constructions were rated remarkably high, patterning much more like the NORMAL sentences. This suggests that this construction may now be the standard. However, the slow response time of participants to the HAVE-got sentences, compared to the normal sentences, suggests that participants may still have some awareness of the social complications of the construction, thus causing longer processing times. The COME sentences had both middling responses and slow RTs. Of all the manipulated sentences in my study then, these may be most presently variable.

The SPEAKER and TYPE of sentence interaction showed that while certain

sentences sound most grammatical with certain speakers, other sentences may sound worst with those same speakers. The interaction was not entirely straight forward, and is complicated by the fact that the participants heard few and different sentences from each of the speakers. But the patterns seem to suggest that speakers who are rated best with the NORMAL and GOT sentences were rated worst with the preterites, and vice versa. The patterning of the BAD sentences was less clear: Speakers K and A were both rated highly with the NORMAL and GOT sentences and poorly with the preterites, but Speaker K was also rated amongst the best with the BAD sentences, while Speaker A was rated the worst.

The differences between the speakers appear to be related to the age and class ratings they received in the SP Experiment, such that the older and more professional sounding speakers were the ones who were rated best with the more standard sentences. However, there will also be an element of each speakers idiolect, the way they performed the sentences, and the comfort with which they read things they would not normally say. In other words, these results are much like those of Walker (2005): who the speaker is appears to affect the GJs, but there are too many variables (prosody, intonation, reading manner) to know if it is due to the social attributes of the speaker alone.

With SPEAKER and TYPE being held constant, we also have an effect of the AVERAGE AGE rating given to each sentence in the SP Experiment. Presumably, this factor captures specific features of the particular recording that altered age ratings, such as the particular sentence or specific production elements, which could conceivably include the realisation of phrase final /t/. What the AGERATING effect tells us is that the older a speaker sounded in a particular sentence, the less grammatical the sentence was rated. This effect was strongest in the COME sentences, which are primarily produced by young and especially non-professional speakers.

At this stage it is worth revisiting some of the assumptions made about the sentences in the methodology section of this chapter. Namely, I assumed that the NORMAL sentences would be socially neutral, because they contained constructions that had not been documented as varying in NZE. However, the sentences were all created by the young author, and thus, unintentionally may have contained turns of phrase and vocabulary that were more likely to be used by younger speakers. Though I am not sure how it could easily

have been done, I did not test the likelihood of a younger or older speaker saying any of my sentences. Thus, its possible that the AGERATING effect worked on the NORMAL sentences (which we must assume because of the lack of an interaction) because the NORMAL sentences in my study, like the COME sentences, are more likely to be produced by younger speakers. In the future I intend to revisit this question by getting age and class ratings of the sentences when presented visually to participants.

The CLASSRATEDIFF effect reinforces the AGERATING effect by showing that the more a released /t/ raised the class rating of a speaker, the less grammatical a sentence would be rated. The immediate question that needs answering is why it was the CLASSRATEDIFF and not the AGERATEDIFF that had this effect. While it was possible to substitute AGERATING with CLASSRATING for a less effective model than the one in Tables 5.3 and 5.4, it was not possible to subsitute AGERATEDIFF for CLASSRATEDIFF in the RATEDIFF model (Tables 5.6 and 5.7). The average age and class ratings given to the sentences were highly correlated (Spearmans Rho=0.8658149, p=<0.001). However, it was the age ratings that were more significantly altered by the phrase final /t/ in the SP Experiment.

There are a number of potential explanations for this CLASSRATEDIFF over AGERATEDIFF effect. In terms of the statistical model, as the released /t/ was more effective in altering the age rating, there may not have been the same variation in AGERATEDIFF values that there was in the CLASSRATEDIFF values, meaning that it was harder for an effect to come out. In the case of the age ratings being so potent, however, we would probably expect a realisation effect in the overall models if this was the only explanation.

Another reason could be that because the realisation of /t/ was less effective in changing class ratings (possibly for the reasons discussed in section 4.3), it was only the strongest and most salient differences between the released and unreleased /t/ that had any effect. These stronger /t/ were then more able to force an effect in the grammaticality ratings (whereas the realisations that had slightly altered age ratings would be less powerful). In a similar vein, but with different implications, the released /t/ that was strong enough to alter the class rating could be so marked and odd that its presence simply made sentences sound less grammatical.

Finally, it could be that the class of the speaker was a more important factor in rating the grammaticality of the sentences than their age. This could be because class was a more important factor in the acceptability of the constructions themselves (i.e., COME has stonger class connotations than age ones), or because people associate grammaticality more with social class than with age.

One final consideration is why the GOT sentences were approaching significance in the Wilcoxon tests comparing responses to the innovative and conservative realisations, while the COME sentences, which were most affected by the AGERATING effect, did not. As mentioned in the summary, the HAVE-got constructions were the only instance where the /t/ was in invariant part of the construction. Therefore we would expect, if the /t/ was to have an effect anywhere, it would be here. And indeed, if the construction had received the sort of variable responses that the COME sentences did, then it is possible the effect of the realisation would have reached significance. However, the HAVE-got sentences were rated as so highly grammatical that they may have plateaued, which is why we do not see an independent AGERATING effect on these sentences. Similarly, if the /t/ had occurred in the invariant part of the COME constructions (namely in the word come), then we may also have seen a significant effect of REALISATION, just as we see strong social effects in AGERATING.

In terms of the hypotheses I put forward at the start of this chapter, these results do not provide a definitive answer, but certainly begin to paint a picture. Phonetic detail was used in this thesis to represent, in a highly controlled fashion, any effect that speaker attributes could have on the grammaticality ratings given to the sentences. What we see in these results is that phonetic detail that saliently carries class information does alter GJs, as do other speaker-related factors. I would say this renders the null hypothesis false.

The direction of speaker-related effects, and the way in which these appeared to interact with the type of sentence, are best explained as being primarily experience-based. Speakers who are younger and non-professional sounding are rated better with the preterite sentences than speakers who are older and professional sounding. With this effect accounted for, there is an additional preference for younger sounding sentences overall, which appears to be strongest in the most socially variable of the sentences, the preterite COME

constructions. We also have a simialar effect of /t/ realisation, such that the more a released /t/ made a speaker sound like they were of a higher social class, the more that /t/ lowered the grammaticality rating of the sentence.

It appears, at the same time, that there may also be an effect of prejudice. In constructions not noted for social variation, sentences with the oldest and most professional sounding speaker (Speaker K) are rated as most grammatical. While we would not expect participants to have encountered these constructions any more from older professionals than from younger nonprofessionals, it seems reasonable to assume that the former group are more associated with grammatical and proper speech, such that, in an absence of experience-related data telling them otherwise, participants will rate older and professional speakers as more grammatical on this association alone.

5.8 Conclusion

The results of the experiment described in this chapter show that speaker attributes can affect grammaticality ratings, and that with this being the case, phonetic detail that is strongly marked for speaker attributes can likewise alter GJs. Such speaker effects show signs of interacting with the type of construction being rated, in a way that suggests participants will rate a socially variable construction better if produced by the type of speaker who they would have encountered using it in the past. The implications of these findings, theoretical and methodological, will be discussed in detail in the next chapter.

Chapter 6

Discussion

Auditorially presented Grammaticality Judgements can be affected by speaker related factors, and even by the realization of a single, socially-salient phonetic variable. The direction of these effects in my data showed that participants appeared to give higher grammaticality ratings when certain speakers used certain constructions. On top of this was an overall preference for younger (or non-professional) speakers, which was strongest in a construction which is used most by younger (and non-professional) speakers. The cause of these effects could be many-fold. The participants could be consulting prejudices or experience, and this information could be meta-linguistic or, as I will argue, grammar internal. Independent of these questions, the methodological implications of my findings are indisputable.

This chapter is split into two main sections. The first (6.1) deals with the theoretical implications of my results, looking at the arguments for and against a grammar-internal explanation, and finishing with a discussion of the phonetic-syntax interface (6.1.3). The second section (6.2) looks at two of the methodological implications of my study, the first for the collection of GJs (6.2.1) and the second for the calculation of frequencies (6.2.2). Finally, in Section 6.3, I discuss further possible research to grow from this study.

6.1 Theoretical Implications

No one would argue that there are methodological implications of the finding that speaker and speaker-related factors affect GJs, but there will undoubt-

edly be those who would argue that there are no theoretical implications, at least none relevant to morpho-syntax. Such an argument would dismiss the effects as products of performance, being extra-linguistic processes outside the grammar, rather than products of competence, being grammar internal.

The alternative argument, of course, is that the effect does reflect competence: I will argue that grammaticality is dependent on speaker, because speaker information is automatically stored with and attached to constructions. I will argue that phonetic detail can affect grammaticality either because of its social associations, or because it is also stored with encountered constructions.

6.1.1 Are Speaker Effects Grammar External?

"...usage, frequency, and so on, are not represented in the grammar itself" (Newmeyer 2003, 682).
I expect that Newmeyer will not feel misrepresented if I include speaker effects, like the ones I have discussed, as belonging under the so on that he refers to in the quote above. This is not to say that the effects arent considered to be real in such an approach, but rather that they are seen as the result of the meta-linguistic processes involved in making grammaticality judgements, and not the internal grammar of language users.

In the process of making a GJ then, informants would consult their internal grammar, which would produce the same answer irregardless of speaker (and other extra-linguistic information). This answer would then go through some sort of world-knowledge filter, which would modify the grammars response - a sort of performance in perception. So, for example, there might be a prescriptivist filter that, regardless of what their internal grammar permits, rejects or is more critical of constructions that are prescriptively ungrammatical.

To explain my results, we could posit that amongst participants extralinguistic processes was a bias towards rating utterances by younger and non-professional speakers as more grammatical. This could be due to an ingroup preference (my participants were generally younger), a priming effect of the young experimenter, or a prejudice that older, professional speakers do not speak ungrammatically (and so the sentences sound less appropriate from

them).

While I do not feel I can rule out any of the above explanations based on my data, my results weakly suggest that the effect, or at least the strength of the effect, that speaker has is dependent on the construction itself. If this is the case, then a meta-linguistic explanation becomes less plausible. If participants have a real world filter that marks this construction as more grammatical with a professional speaker, but that construction as less grammatical with a professional speaker, then isnt that speaker information and that construction directly associated with each other? And if they are, how is that any less part of the grammar than other directly associated information, such as animacy, tense, semantic class or prosody? Where do we draw the line about what effects we do and dont allow in our grammar?

6.1.2 Are Speaker Effects Grammar Internal?

Could it be the case that, like traditionally linguistic factors, speaker information affects grammaticality judgements because it is part of the grammar? That is, the internal grammar itself varies with the speaker. In this sense, language users would not have a single, independent and strict grammar in their heads. Rather, they would have access to either a finite number of self-contained grammars that reflect the different speaker groups they encounter, or to a single but fluid and continuous grammar, with fuzzy rules and probabilities that are dependent, amongst other factors, on who the speaker is.

That language users carry different grammars for different dialect groups is a seductive solution. It would allow speaker information to have an effect on GJs, but only because a speaker of a certain dialect or sociolect causes the independent grammar of that dialect or sociolect to be accessed, just like, when listening to someone speaking Māori, I would access my Māori grammar over my English one. Such a view need not impact on the notion that grammars, as a phenomenon, are categorical and context-free, because the speaker information is only involved at the level of which grammar is selected, but not in the online processing of the rules of that grammar.

This sort of account is problematic on at least two counts. The first is the question of how many grammars there are, and on what speaker distinc-

tions we make these grammars. While it might be quite plausible to say that I have, for example, a grammar for speakers from Southland, and a different grammar for the rest of New Zealanders, it becomes more difficult to argue that I have a grammar for young speakers and old speakers and male speakers and female speakers and non-professional speakers and professional speakers and Māori speakers and Pāakehā speakers. Furthermore, if I encounter a young, non-professional Māori speaker, which grammar do I access? The grammar I have for young speakers, or for non-professionals or for Māori? Or the grammar I have for young, non-professional Māori speakers, which would lose the similarities this might have with, say, and older, non-professional Māori speaker. And how young does a speaker have to be to invoke the young grammar? How old to invoke the old grammar?

The second, and I think much more important problem, is that my effects appear to be gradient and continuous. Phrase final /t/ realizations that altered class ratings also altered grammaticality ratings, but it would be odd to assume that this realization of the /t/ caused participants to switch from using their non-professional to their professional grammar. Similarly, the speaker effects did not appear to make sentences jump from being grammatical to non-grammatical, but rather made them a little more, or a little less, grammatical.

Thus, I think the best explanation for my data (though my data is not the best possible evidence for this explanation), is that speaker information is incorporated inside a single grammar which is fluid and continuous. While such detail has long been considered part of performance, in his argument for including probabilities in grammars, Manning notes that over the last few decades "the scope of grammar has been expanded in various ways: people now routinely put into grammar semantic and discourse facts that would have been excluded in earlier decades" (Manning 2003, 307). While tradition would have us dismiss this data, traditions change. Moreover, tradition should not drive our explanations; results should.

6.1.3 An Exemplar Model of Syntax

"*Syntactic investigation of a given language has as its goal the construction of a grammar that can be viewed as a device of some sort for producing the*

sentences of the language...linguists must be concerned with the problem of determining the fundamental underlying properties of successful grammars" (Chomsky 1957, 1).

While most present models of syntax would struggle to include speaker information as an underlying property of the grammar, an exemplar model, like those used in phonology (see 2.1), would inherently record salient speaker information with encountered instructions. So, if the use of the preterite form come in a past tense construction is encountered more often coming from younger non-professionals, it will be predicted to be rated as more grammatical when said by younger non-professionals. This is based on the assumption that the grammaticality rating of a construction, at some level, reflects the robustness of the construction (in combination with its context) in the memory.

My data would support this theory if there had been an interaction, at all stages, of speaker related factors and the type of construction in question. Some of the constructions included (the COME, DONE and HAVE-got sentences) were documented as occurring more in the speech of younger, nonprofessionals. We would expect these constructions to be rated as most grammatical when presented with younger, non-professional speaker information. The rest of the constructions were included with the assumption that they did not vary in a socially meaningful way in production, either because all speakers would be as likely (normal sentences) or as unlikely (bad sentences) to say them. We would expect there to be no speaker-related effects on the ratings of these constructions.

The only interaction between a speaker-related effect and the type of construction was an overall one that showed that the five speakers received significantly different GJs for the same types of sentence, but the hierarchy of these ratings depended on which construction was being rated. For example, while NORMAL, GOT and BAD sentences with Speaker K received the highest grammaticality ratings, the preterite sentences read by her received the lowest ratings. Preterite sentences with Speaker E, meanwhile, received the highest ratings,but the other sentences with her received amongst the lowest. An eyeball of this interaction suggests that it is related to the perceived age and class of these speakers, such that the higher a speakers perceived age and social class, the better they did with the non-variable sentences (I include HAVE-got amongst these), and the worse they did with the preterites.

However, as this effect was not the focus of the experiment, a number of confounding factors were not controlled for (the five speakers were reading different sentences), and there is no statistical verification of this trend. In other words, this interaction is generally supportive of an exemplar model, but could also easily be dismissed as too fragile.

Other speaker related effects we saw were an overall improvement in grammaticality ratings the younger or lower the class rating that sentence had received in the SP Experiment, and similarly an overall effect of the realization of the phrase final /t/, such that the more the unreleased variant had garnered a lower class rating relative to the released variant, the more the sentence was rated as grammatical with the unreleased variants. The direction of these effects is what an exemplar model would predict for the target sentences, but not for the NORMAL and BAD controls, where we would expect no effect. This would seem to suggest that the effect is based not on experience, but on a bias or prejudice against older sounding sentences (though the generally high ratings of Speaker K contradict this). However, as I discussed at the end of the last chapter, assuming that the NORMAL and BAD sentences were socially neutral was quite probably a flawed assumption, and these sentences may in fact have been more likely to come from younger speakers. It should also be noted that the AGERATING effect was strongest in the COME sentences, which are possibly the most socially variable of all the constructions[1].

One more result that is important in this discussion comes from the SP Experiment. Though participants were told not to pay attention to what was being said in rating the age and class of participants, the type of construction being rated still affected the ratings they gave to the speaker, in a way that mostly reflected production. That is, if the speaker was saying a preterite sentence as opposed to a NORMAL sentence, they would be perceived as being younger and of a lower social class. This would support an exemplar model, because those constructions would be stored with such speaker information. The complicating factor in these results was the behaviour of the BAD sentences, which should have no speaker associations, and thus, we

[1]The ratings given to the HAVE-got sentences were almost as high as the NORMAL sentences, and the response time to the DONE sentences was almost as quick as the BAD and NORMAL sentences, which all suggests that these constructions are less variable (GOT always good, DONE always bad) than the COME sentences.

might expect them to be rated somewhere between the ratings given to the preterites and the NORMAL and GOT sentences. This is the case for the class ratings but not for the age ratings, where they are rated lowest. Thus, there could be some sort of prejudice that states that ungrammatical sentences are said mostly by young speakers, and this alone explains the effect. However, it could also simply mean that, being something that people never encounter, the social associations of BAD sentences are complicated.

The argument for an exemplar model of syntax, then, where encountered constructions are stored complete with information about who says them, can not be made using my data alone. However, I believe that my data adds to the growing literature that suggests it is an avenue we should pursue further.

6.1.4 Phonetic Effects on Grammaticality Judgements

Gahl and Garnsey (2004), and Hay and Bresnan (2006) show that phonetic detail is affected by syntactic factors, and my study shows that the grammaticality ratings of morpho-syntactic constructions are affected by phonetic detail. The socially salient realization of a phrase final /t/ affected the grammaticality ratings given to a sentence.

In the previous section I discussed how an exemplar model of syntax could explain some of my results. Such models assume that encountered instances of speech are stored as complete memories, including speaker information and the phonetic signal. This type of account has successfully explained perceptual effects seen in phonology, where altering speaker information can alter the word people report hearing (refer to the review in 2.1.1). It also successfully explains the results of the SP Experiment I describe in this thesis, where altering the signal can alter who people believe they are listening to.

Thus, if we were to make an argument for ET using my results, we might want to say that the reason that the constructions were rated as more grammatical with an innovative realization of the phrase final /t/ (when such a realization caused the speaker to be rated as of a lower class) was simply because participants had encountered more instances of come as a simple

past tense verb in signals that include an unreleased phrase final /t/ than in signals that include a released phrase final /t/. This would be because each construction had been stored with the complete phonetic signal from each encounter, and the stored signals were being accessed.

An alternative explanation is that while constructions may be abstracted from detailed phonetic memories, and still weakly attached to specific memories, the effect of the /t/ comes from its association with speaker information, and not because of a direct link with the construction. While we may store every the phonetic signal of every sentence we encounter, the memory of this phrase will decay over time in the absence of further reinforcement (Hay & Bresnan 2006, 346). Similarly, even if the signal is stored for a long time, the phonetic signals of encountered preterite come constructions will vary wildly, just because the words in and around the constructions can vary widly. So for the hidden structure of a preterite come construction, we would expect weak or nearly non-existent memories of the phonetic detail, except potentially in the invariable word come. However, the speaker attributes of stored instances of preterite come constructions would presumably vary less than the phonetic detail, and, for example, the label of non-professional will be tagged in the majority of such exemplars. A released phrase final /t/, regardless of construction, has strong links with professional speakers (as seen in the SP Experiment). So hearing a released /t/ with a preterite come construction might be rated poorly, not because you havent heard many released /t/ with the construction, but because the /t/ reminds you of professionals, and you havent heard many professionals using the construction.

It is impossible from my data to know whether it is the storage of the /t/ with the construction , or the association of the /t/ with the speaker that is responsible for the effect. The fact that the only almost significant effect of /t/ realisation overall in the Wilcoxon tests was in the GOT sentences, where the phrase final /t/ occurs in the invariant part of the construction, suggests an effect of phonetic storage. The fact that it was only the very socially salient /t/ that were able to alter the ratings suggest an effect of the social associations of the phonetic detail. I imagine that it is the case that both mechanisms are at work, and ET can certainly account for both mechanisms being at work.

My results and others suggest that there is an interface between phonetics

and syntax, and that by looking at the two together we can learn new insights about both and about language use in general. They do not appear to be wholly separate and independent systems, which suggest that perhaps there are no wholly separate and independent systems in language at all.

6.2 Methodological Implications

Irrespective of potential disagreements over the theoretical implications of my data, the results presented in the previous chapter have indisputable and serious methodological implications. In this section I first look at what my results mean for the collection of GJs, and then at what they could mean for the calculating of frequencies and probabilities.

6.2.1 Collection of GJs

Whether based on linguistic internal or external processes, if informants are sensitive to speaker information, this is an additional element that needs to be controlled for in any auditory presentation of stimuli, which is a method advocated by Kitagawa and Fodor (2006). Having different speakers for different constructions could complicate results, and differences between studies could be due to differences between the speakers. More alarmingly, the results of my study suggest that even with the same speaker, different realizations of socially salient phonemes may also affect judgements.

The simple response to these findings, then, would be to continue with standard practice and present stimuli in written form. This is still the most common, and perhaps my data would suggest, the safest course of action.

An alternative interpretation, however, would be that these results show we need to always present stimuli in spoken form. The reason for Kitagawa and Fodors endorsement of aurally presented stimuli was their valid concern that acceptability judgements on written sentences are not purely syntaxdriven; they are not free of prosody even though no prosody is present in the stimulus (Kitagawa & Fodor 2006, 358). Participants appear to be silently reading sentences with a default prosody, which affects their judgements of prosodically-sensitive constructions. Since we cannot guarantee which prosodic pattern participants will use, inter and even intra participant variation could be due

to prosodic variation amongst informants, which leads the authors to say: prosody needs to be under the control of the linguist who solicits syntactic judgements, not left to the imagination of those who are giving the judgements (Kitagawa & Fodor 2006, 358). Aurally presented data is a means for the linguist to control this variable.

Much like Fodors Implicit Prosody Hypothesis (Fodor 2001), where in silent reading a default prosody is projected onto the stimulus, it seems reasonable that I propose here an Implicit Speaker Hypothesis: in silent reading, a default speaker is projected onto the stimulus. As expressed by the poet (Lux 1997, 15):

> *THE VOICE YOU HEAR*
> *WHEN YOU READ SILENTLY*
> *is not silent, it is a speaking*
> *out-loud voice in your head: it is spoken,*
> *a voice is saying it*
> *as you read. (1-6)*

Experiments by Alexander and Nygaard (2008) found that participants who had been familiarised with a fast and a slow speaker read passages aloud and silently faster when they were told that it had been written by the fast speaker, suggesting that readers engage in a type of auditory imagery while reading that preserves the perceptual details of an authors voice (Alexander & Nygaard 2008, 446). When we ask participants to read sentences where they do not know the author, we cannot assume that participants read sentences with speaker neutralized, nor can we, as is probably more commonly believed, assume that participants read the sentences with themselves as speaker. They well might, but they could also be projecting, for example, an RP speaker on to the stimuli. They could even be altering which speaker they project onto a sentence depending on which sort of speaker would be most likely to say a sentence, which is one of the potential strategies described by Manning: "humans judge the grammatical acceptability of sentences by assuming a most favorable real world context" (Manning 2003, 310).

My data suggests that who says a sentence affects the grammaticality judgement given to that sentence. While the Implicit Speaker Hypothesis certainly

requires some explicit testing, until we better understand if and who participants hear saying visually presented constructions, then the safest option maybe well be to impose a speaker on them by presenting stimuli auditorally. How we wish to do this will vary with what our question is. It may be in the researchers best interest to have one speaker of the language for all constructions, or it may pay to try out each construction with a range of speakers.

6.2.2 Probability and Frequency Estimations

The speaker-induced differences in my grammaticality ratings appear at some level to reflect frequency. Certain speaker use certain constructions more, and thus those constructions are rated as more grammatical when said by those speakers, as opposed to by groups who use the construction less. However, as I noted in Chapter 2, the literature to date suggests no clear relationship between frequency and perceived grammaticality. While I believe that this is probably due to the fact that GJs are complex performances that dont entirely reflect the grammar (and of course, could just be because GJs do not reflect frequency), the lack of conclusive evidence may also be due to the overly simplistic and non-representative way that frequencies are generally calculated.

The reason to include frequency counts in studies is that the number of times a participant has encountered a word/speaker/construction clearly has an effect on both production and perception. Frequencies are generally calculated as logarithmic counts across corpora, which can either be written or spoken. One of the most commonly used sources for English words is the CELEX Lexical Database (Baayen, Piepenbrock & Gulikers 1995), which uses the 17.9 million word Cobuild corpus (Renouf 1987), which consists of primarily British but also some American written texts. Usually frequencies are calculated on one or two dimensions. For example, Manning and Schutzes (1999) probabilistic context free grammar (PCFG) and variations on it focus onlexical and syntactic information in estimating the probability of a parse structure (Crocker & Keller 2006, 237).

While these are fine as rough and general guides, these sorts of calculations simplify the sort of frequency counts that language users have access to. If

we are indeed storing not just a construction, but a whole memory, then the frequency of the accompanying details of the memory also become important, and thus could also affect GJs. My results are evidence of this: it was not just the construction, but the accompanying information about who used the construction that affected how people rated the sentences. Therefore, in calculating the frequencies of constructions (or words), we should be attentive to what kind of speakers we are calculating our frequencies over, and indeed, should include speaker attributes as a contributing dimension in frequency counts. A null correlation, otherwise, does not necessarily mean that GJs do not reflect frequency, but simply that we are not correctly calculating the relevant frequencies for our participants, for the token at hand.

A number of recent studies outside of syntax have shown that more sophisticated participant and/or token specific treatments of frequency can prove insightful. In their study of t/d deletion in running speech, Guy, Hay and Walker (2008) found that the frequency of a word did not affect the likelihood of deletion if the frequency was taken from the CELEX database, but did affect it (such that more frequent words were more likely to undergo deletion) if frequency was calculated locally, over the speakers they used in their study. Considering that speakers of the CELEX database and speakers of their study differed substantially in both dialect and age, their finding was not surprising, but did highlight the danger in using frequency counts that werent relevant to the subjects in a study.

In their study on phrase final /t/ in NZE, Docherty et al. (2006) found not only that frequent words were produced with the unreleased variant more than infrequent words, but that words that were more frequently in a phrase final position also showed more unreleased variants than words that were infrequently in the same position. A basic frequency count of words would capture the first fact, but not the second, which is more token specific. In the Guy et al study, it was not only the following environment of the token in question that affected whether the final consonant was deleted, but also the single most comon following environment that usually followed the word in question, as predicted by Bybee (2003). That is, whatever online, environmentally-driven processes might affect the deletion of /t,d/ in production, the product affects the representation of the word in peoples heads, such that a frequent following environment can affect the realization of the final consonant in a word, even if the particular instance doesnt have that

following environment. Guy et al incorporated the frequency of the following environment in their analysis, much like their speakers appeared to.

Returning to GJs, speakers may be using frequency in their ratings, but may be calculating the frequencies on very token specific information: How frequent is this construction in the specific frame? How frequent is this construction with these words? How frequent is this construction from this speaker? As Bresnan and Nikitina (2003) found, various factors influenced the probability of which dative alternation subjects used and predicted others had used, and noted that examples that have been reported in the literature as ungrammatical "tend to utilize the far less frequent positionings of argument types" (Bresnan 2005, 12). Thus, we may not be finding a correlation between grammaticality and frequency simply because, while we may be looking at a frequent or infrequent construction, we may be presenting it with overly infrequent or frequent corresponding contextual information.

My own study would have been improved if I had factored in the probability of each sentence I had participants rate in terms of not only the construction, but attributes of the speaker and other contextual information (specific words, topic, the specific frame, etc). This, of course, would require a complex, annotated corpus over which to calculate frequency, and by parsing the numbers on so many dimensions, the numbers would get too small for a reliable count unless the corpus was particularly large. This may mean that we have to choose fewer dimensions that production data has suggested will be particularly influential.

Sweeping and simple frequency counts are not without merit, but the sums do simplify what is actually a multi-dimensional phenomenon. Frequency is instrinsically multi-dimensional because encountered instances of speech are multi-dimensional. For every salient feature that language users attach to a construction, or word, or phoneme, our models will improve if we alter our frequency calculations to incorporate it. I believe that my results have shown that speaker information is one such feature, and to ensure the robustness of our frequency calculations we may only want to include it when we have good reason to believe it will be influential.

6.3 Higher Up and Further In

This thesis could be followed by a number of related and expanding experiments.

The results of my SP Experiment show that the realization of a single phonetic variant can significantly alter the perceived age or class of a speaker. My preliminary investigations suggested that while this worked for phrasefinal /t/, the manipulation of other variables was less successful. Discovering the extent to which different variables can alter judgments and trying to posit the reasons behind this could be a valuable course of research. Leading on from this, it would also be interesting to see the effect of manipulating more than one variable in a sentence. In Walker (2007), I did this and found that two variables manipulated in a socially aligned manner resulted in stronger age and class manipulations than either did on their own. One could also test the effect of putting two non-aligned variables in a sentence. The results of my GJ Experiment were unclear enough to render a similar, but improved, study highly worthwhile. In such a study I would draw the inspiration for my sentences from real speech, changing only what is necessary for experiment design and control. I would also make sure I knew the likelihood of each sentence, including the controls, in terms of a number of factors, especially speaker information. I also think that a Magnitude Estimation Scale, as opposed to a six point one, might lead to more interpretable results, and I would be tempted to test the manipulation more than one variable in the sentence.

Following on from the work of Gahl and Garnsey (2004), and Hay and Bresnan (2006), which showed that syntactic features affect phonetic realizations, we could run another GJ elicitation task that tested the ability of phonetic detail to alter GJs without reference to the social associations of such detail. For example, Hay and Bresnan showed that the nucleus of hand was more raised when referring to the limb than when in expressions such as give a hand. Thus, it might be possible that a raised realisation of hand would elicit higher grammaticality ratings with its literal meaning than would a lowered realisation, and than it would with its metaphorical meaning.

There are other ways to ascertain whether speaker information is stored in the grammar. A number of studies (Strand & Johnson 1996, Drager 2005, Hay,

Warren & Drager 2006) used photo manipulations to alter the perception of an incoming signal. My study could similarly be recreated with a photo manipulation instead of the phonetic one, such that the same socially marked sentence is presented twice to participants, one time with a photo/video of a younger speaker, another time with a photo/video of an older speaker.

In this experiment I tried to prime different sociolects, but the same study could be run, and could possibly be more successful, looking at either ethnolects or dialects. AAVE, for example, has a range of morpho-syntactic constructions (Labov 1998) that do not occur in SAE. Thus, the constructions might be rated as more grammatical when presented in conjunction with phonetic variants of AAVE or with a photo of a black speaker, than if presented with variants of General AmE or a photo of a white speaker. Similarly, priming participants to the concept of Southland as opposed to the concept of New Zealand in general might result in higher grammaticality ratings of constructions like *The baby needs fed*, which are a feature of Southland English.

Finally, the Implicit Speaker Hypothesis that I propose in this chapter demands further investigation. This thesis provides no empirical evidence for the theory, but my data highlights the problems with written presentation of stimuli, at least until we better understand who people hear when they silent read. The nature of silent reading makes the task of collecting empirical evidence for the hypothesis particularly difficuly, but the study of Alexander and Nygaard (2008) offers a promising template. In particular, researching whether readers alter the voices they hear relative to the type of sentence that they are reading seems pivotal.

6.4 Conclusion

My results have both theoretical and methodological implications for the field of morpho-syntax. The theoretical implications are to some extent dependent on how one interprets my results, but I argue that generally they suggest that speaker information is stored as part of the grammar, which is best explained by extending ET to morpho-syntax. Such an account of grammar greatly challenges traditional accounts, and my conclusions need more more empirical support, but this is what we are here to do.

There can be no argument over the methodological implications my results have for the collection of grammaticality judgements. However, while the demonstrated effect that speaker and phonetic detail can have on GJs may deter some fieldworkers from using auditory presentations of stimuli, I argue that unless we know who (if anyone) informants hear when reading sentences, we may in fact be better to always present the stimuli auditorially so at least any potential speaker effects are controlled by the experimenter. I also argue that the calculations of frequencies and probabilities should take more factors into consideration, including speaker information.

Chapter 7

Conclusion

The motivation for this thesis was to find evidence that grammatical constructions are stored with speaker information and possibly phonetic detail. Such evidence would support an exemplar model of syntax, building on from the usage-based models being developed in syntax and expanding the domain of ET, which has been considered primarily in regards to phonology. To test this hypothesis, I designed and implemented an experiment to find out whether the manipulation of a single, socially-salient phonetic variable could alter the grammaticality judgements given to socially variable morphosyntactic constructions.

So that I had a sufficiently socially salient variable for the main experiment, I ran a Pilot Study in which I tested the ability of a range of documented phonetic variables in NZE to affect the age and social class ratings given to speakers. The results suggested the manipulation of phrase final /t/ was best able to change perceived speaker attributes. The fact that the other variables were not able to change the ratings could reflect aspects of the experiment design, or may reflect differing levels of either aural or social saliency, or both.

Deciding to focus on phrase final /t/, I ran another speaker perception test that tested the variable in the same sentences, and indeed, the same recordings that were to be used in the grammaticality judgement experiment. In this more controlled experiment, the /t/ more significantly altered both the age and class ratings given to the speakers, such that when a sentence was presented with a conservative realisation of the variable, the speaker was rated as being older and of a higher social class than when the same sentence

was presented with the innovative variant. The fact that this mirrors the production patterns that language users encounter shows that such phonetic detail must be stored with speaker information. The fact that the detail is able to alter perceived speaker attributes in a sentence filled with other information shows the sensitivity of listeners to very fine phonetic detail. Another important finding of the Speaker Perception experiment was that, despite the fact that participants were told to not pay attention to what was being said in rating speaker information, the type of construction in the sentence also affected the age and class ratings given to the speakers. Again, the way in which the effect worked reflected patterns seen in production, such that the speakers were rated as younger and of a lower social class in the sentences with non-standard constructions, which are used mostly by younger non-professionals. This is indication in itself that encountered constructions are stored with speaker information.

The same recordings were then used in the GJ Experiment, with various results from the Speaker Perception experiment tested in the logistic regression model as independent predictors. The type of construction being rated unsurprisingly affected the ratings, such that the less standard constructions were rated worse than the standard constructions, and the non-grammatical BAD sentences received the lowest ratings of all. The possessive HAVE-got constructions, which had been previously been reported as an incoming variant, and included in the study as a socially-variable construction, patterned very closely with the NORMAL constructions in terms of responses (though not response times) suggesting that the construction is now the standard.

There were a number of socially conditioned effects on the constructions. Which of the five speakers read the sentence affected grammaticality ratings, but this was in an interaction with the sentence type: Speaker K, who was rated as oldest and of the highest social class in the SP Experiment, was rated best with the bad and standard constructions, and worst with the preterite constructions. This would appear to reflect our participants experience of these constructions.

This effect taken into account, there was also an effect of the age rating given to a particular sentence in the earlier thing. This saw sentences that had been rated as having an speaker who was older or of a higher social class being rated as less grammatical. The overall bias in the sentences was

towards being produced by younger, non-professional speakers. This result then, where participants rate sentences as more grammatical from younger, nonprofessional speakers when the sentences have a higher probability of being used by younger, non-professional speakers, may also reflect participants experience.

To specifically isolate any effect that the realisation of /t/ might be having I factored RATEDIFF into modelm being the difference between the ratings given to sentences with a conservative and innovative realisation of phrase-final /t/. I found an effect of CLASSRATEDIFF: The more a conservative version of the /t/ had received a higher class rating than the innovative version in the SP Experiment, the more the sentences was rated as grammatical with the innovative variant. This seems to pattern with the effect discussed above, where participants rated sentences where the speaker sounded younger as more professional. It also suggested that only strongly socially salient phonetic detail can alter grammaticality ratings.

The effect that the social factors had, which seemed to reflect some sensitivity to the patterns seen in production, suggest that morphosyntactic constructions are stored with speaker information, as an exemplar model of syntax would suggest. The fact that phonetic detail also had an effect could suggest that constructions are also stored with fine-grained acoustic signals, though it could also be due to the social associations of the /t/. Exemplar Theory can accomodate either, and indeed both, explanations.

The methodological implications of my study are non trivial. Speaker has an effect, and needs to be controlled for in grammaticality ratings. Specific phonetic realisations can have an effect, and need to be controlled for in the presentation of grammaticality ratings. While this might be a deterrent for many to aurally present GJ stimuli to participants, I argue, suggesting an Implicit Speaker Hyptothesis, that until we know who participants hear in their heads when silent reading, we would be better to orally present stimuli so speaker is at least in our control.

The final methodological argument I make concerns the calculation of frequencies. These are generally done along one or two dimensions, but if language users take multiple factors into account when producing and processing language, then we need to as well. This will not always be possible,

but when we are able to incorporate finer calculations of frequency into our models, it will give us a richer understanding of frequency effects.

This thesis has contributed to the literature by explicitly testing whether listeners use a socially-variable phonetic variable in attributing social values to speakers, and in making grammaticality judgements, and showing that they do. These results provide support in favour for expanding the range of Exemplar Theory to syntax. In this thesis I have also highlighted methodological issues in the aural presentation of GJ stimuli, and in the calculations of frequencies. I also posit an Implicit Speaker Hypothesis, where read sentences are influenced by an effect of a default speaker. Overall, my study shows the positive effects of using phonetics to gain insights into other areas of linguistics that are traditionally kept separate.

Bibliography

[1] Aarts, B. (2007), *Syntactic gradience: the nature of grammatical indeterminacy*, Oxford University Press, Oxford.

[2] Aarts, B., Denison, D., Keizer, E. & Popova, G., eds (2004), *Fuzzy grammar: A reader*, Oxford University Press, Oxford.

[3] Alexander, J. D. & Nygaard, L. C. (2008), 'Reading voices and hearing text: talker-specific auditory imagery in reading', *Journal of experimental psychology. Human perception and performance* **34**(2), 446 59.

[4] Allan, W. S. & Starks, D. (2000), No one sounds like us? a comparison of New Zealand and other southern hemisphere Englishes, in A. Bell & K. Kuiper, eds, *New Zealand English*, Benjamins, Amsterdam, pp. 53 83.

[5] Angermeyer, P. S. & Singler, J. V. (2003), The case for politeness: Pronoun variation in co-ordinate NPs in object position in English, *Language Variation and Change* **15**, 171209.

[6] Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995), *The CELEX lexical database* (cd-rom).

[7] Bard, E., Robertson, D. & Sorace, A. (1996), Magnitude estimation of linguistic acceptability, *Language* **72**, 32 68.

[8] Batterham, M. (1995), *There is another type here: Some front vowel varieties in New Zealand English*, PhD thesis, La Trobe.

[9] Bauer, L. (1992), The second great vowel shift revisited, *English Worldwide* **13**, 253 268.

[10] Bauer, L. (1994), English in New Zealand, in R. Burchfield, ed., *English in Britain and Overseas Origins and Development*, Vol. 5 of The Cambridge History of the English Language, Cambridge University Press, Cambridge, pp. 382429.

[11] Bell, A. (1997), The phonetics of fish and chips in New Zealand: Marking national and ethnic identities, *English World-Wide* **18**(2), 24370.

[12] Bell, A. (1999), The NZ English short front vowels: A conspectus and some data, in *The 10th International Conference on Methods in Dialectology*, St. Johns Newfoundland.

[13] Bever, T. G. & Carroll, J. M. (1981), On some continuous properties in language, in T. Myers, J. Laver & J. Anderson, eds, *The cognitive representation of speech*, North-Holland, Amsterdam, pp. 225 234.

[14] Bod, R. (2000), The storage and computation of three-word sentences. Paper presented at AMLaP 2000, Leiden.

[15] Bod, R. (2006), Exemplar-based syntax: How to get productivity from examples, *The Linguistic Review* **23**, 275 290.

[16] Bresnan, J. (2005), Is syntactic knowledge probabilistic? Experiments with the English dative alternation, in S. Featherston & W. Sternefeld, eds, *The proceedings of the International Conference on Linguistic Evidence, Roots: Linguistics in search of its evidential base*, Studies in Generative Grammar, Mouton de Gruyter, Berlin.

[17] Bresnan, J. & Nikitina, T. (2003), The gradience of the dative alternation, in L. Uyechi & L. H. Wee, eds, *Reality Exploration and Discovery: Pattern Interaction in Language and Life*, CSLI Publications, Stanford.

[18] Bresnan, J., Cueni, A., Nikitina, T. & Baayen, H. (2005), Predicting the dative alternation, in G. Boume, I. Kraemer & J. Zwarts, eds, *Cognitive Foundations of Interpretation*, Royal Netherlands Academy of Science, Amsterdam.

[19] Bybee, J. (1994), Productivity, regularity and fusion: How language use affects the lexicon, in R. Singh & R. Desrochers, eds, *Trubetzkoys orphan: Proceedings of the Montreal roundtable Morphonology: Contemporary responses*, Benjamins, Amsterdam, pp. 247 269.

[20] Bybee, J. (2003), Mechanisms of change in grammaticization: the role of frequency, in R. Janda & B. Joseph, eds, *Handbook of Historical Linguistics*, Blackwell, Oxford, pp. 602623.

[21] Bybee, J. (2006), From usage to grammar: the minds response to repetition, *Language* **82**(4), 711 733.

[22] Campbell-Kibler, K. (2007), Accent, (ING) and the social logic of listener perceptions, *American Speech* **82**(1), 32 64.

[23] Chomsky, N. (1957), *Syntactic Structures*, Mouton co., The Hague.

[24] Chomsky, N. (1965), *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

[25] Chomsky, N. (1995), *The Minimalist Program*, MIT Press, Cambridge, MA.

[26] Cowart, W. (1997), *Experimental Syntax: Applying Objective Methods to Sentence Judgements*, Thousand Oaks, Sage.

[27] Crocker, M. W. & Keller, F. (2006), Probabilistic grammars as models of gradience in language processing, in G. Fanselow, C. Fery, M. Schlesewsky & R. Vogel, eds, *Gradience in Grammar: Generative Perspectives*, Oxford, Oxford University Press.

[28] Davis, P. McLeod, K., Ransom, M., & P. Ongley (1997), *The New Zealand Socio-economic Index of Occupational Status (NZSEI) Research Report 2*, Statistics New Zealand, Wellington.

[29] Davis, P., Jenkin, G., & P. Coope. (2003), *New Zealand Socio-economic Index 1996. An update and revision of the New Zealand Socio-economic Index of Occupational Status. Research Report 20*, Statistics New Zealand, Wellington.

[30] Docherty, G., Hay, J. & Walker, A. (2006), Sociophonetic patterning of phrase-final /t/ in New Zealand English, in P. Warren & C. Watson, eds, *Proceedings of the 11th Australasian International Conference on Speech Science & Technology*, University of Auckland, New Zealand.

[31] Drager, K. (2005), *The influence of social characteristics on speech perception, Masters thesis*, University of Canterbury, Christchurch.

[32] Durkin, M. E. (1972), *A study of pronunciation, oral grammar and vocabulary of west coast schoolchildren*, Masters thesis, University of Canterbury, Christchurch.

[33] Elliot, D., Legum, S. & Annear Thompson, S. (1969), Syntactic variation as linguistic data, in R. I. Binnick, A. Davison, G. M. Green & J. L. Morgan, eds, *Papers from the fifth regional meeting of the Chicago Linguistic Society*, Chicago Linguistic Society, Chicago, pp. 52 59.

[34] Fanselow, G., Fery, C., Schlesewsky, M. & Vogel, R., eds (2006), *Gradience in grammar: Generative perspectives*, Oxford, Oxford University Press.

[35] Fodor, J. D. (2001), Prosodic disambiguation in silent reading, *NELS* **32**, 113 132.

[36] Ford, M., Bresnan, J. & Kaplan, R. (1982), A competence-based theory of syntactic closure, in J. Bresnan, ed., *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge, MA.

[37] Fromont, R. & Hay, J. (2008), Onze miner: Development of a browser-based research tool, *Corpora*, **3**, 173-193.

[38] Gahl, S. & Garnsey, S. M. (2004), Knowledge of grammar includes knowledge of syntactic probabilities, *Language* **82**(2), 405410.

[39] Gleitman, L. R. & Gleitman, H. (1970), *Phrase and paraphrase: Some innovative uses of language*, W. W. Norton and Company, New York.

[40] Goldinger, S. D. (1996), Words and voices: Episodic traces in spoken word identification and recognition memory, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **22**, 11661183.

[41] Gordon, E. M. (1997), Sex, speech and stereotypes: Why women use prestige forms more than men, *Language in Society* **26**, 47 63.

[42] Guy, G., Hay, J. & Walker, A. (2008), Phonological, lexical, and frequency factors in coronal stop deletion in early New Zealand English. Poster presented at Laboratory Phonology 11, Wellington.

[43] Hawkins, J. A. (2004), *Efficiency and Complexity in Grammars*, Oxford University Press, Oxford.

[44] Hay, J. & Bresnan, J. (2006), Spoken syntax: The phonetics of giving a hand in New Zealand English, *The Linguistic Review* **23**(3), 321 349.

[45] Hay, J. & Drager, K. (forthcoming), Stuffed toys and speech perception. To appear in *Linguistics*

[46] Hay, J. & Maclagan, M. (forthcoming), Social and phonetic conditioners on the frequency and degree of intrusive /r/ in New Zealand English. To appear in D. Preston & N. Niedzielski, eds., *Methods in Sociophonetics*, Mouton de Gruyter, New York.

[47] Hay, J. & Sudbury, A. (2005), How rhoticity became /r/-sandhi, *Language* **81**(4), 799 823.

[48] Hay, J., Drager, K. & Warren, P. (2006), Cross-dialectal exemplar priming. Poster presented at Laboratory Phonology 10, Paris.

[49] ay, J., Nolan, A. & Drager, K. (2006), From fush to feesh: Exemplar priming in speech perception, *The Linguistic Review* **23**(3), 351379.

[50] Hay, J., Walker, A. & Drager, K. (under review), Getting your facts right: The effect of affect on speech perception and production.

[51] Hay, J., Warren, P. & Drager, K. D. (2006), Factors influencing speech perception in the context of a merger-in-progress, *Journal of Phonetics* **34**(4), 458 – 484.

[52] Heide, D. (2002), *Grammaticality and acceptability in linguistic research: Refinements in the elicitation and evaluation of judgment data*, MA Thesis, University of Wuppertal, Wuppertal.

hintz Hintzman, D. L. (1986), 'schema abstraction' in a multiple-trace memory model, *Psychological Review* **93**, 328 – 338.

hirst Hirst, G. (1981), *Anaphora in natural language understanding: A survey*, Springer-Verlag, Berlin.

[53] Holmes, J., Bell, A. & Boyce, M. (1991), *Variation and change in New Zealand English: A social dialect investigation*, Technical report, Social Sciences Committee of the Foundation for Research, Science and Technology, Victoria University, Wellington.

[54] Hooper, J. (1976), Word frequency in lexical diffusion and the source of morphophonological change, in W. Christie, ed., *Current Progress in Historical Linguistics*, North Holland, Amsterdam, 96 – 105.

[55] Johnson, K. (1997), Speech perception without speaker normalization: An exemplar model, in Johnson & Mullennix, eds, *Talker Variability in Speech Processing*, Academic Press, San Diego, 145 – 165.

[56] Johnson, K. (2001), Spoken language variability: Implications for modeling speech perception, in R. Smits, J. Kingston, T. M. Nearey & R. Zondervan, eds, *Proceedings of the Workshop on Speech Recognition as Pattern Classification (SPRAAC)*, Nijmegen: Max Planck Institute for Psycholinguistics.

[57] Johnson, K., Strand, E. & DImperio, M. (1999), Auditory-visual integration of talker gender in vowel perception, *Journal of Phonetics* **27**, 359 – 384.

[58] Joos, M. (1966), Description of language design, reprinted in Aarts et al. (2004), 349 – 356.

[59] Kitagawa, Y. & Fodor, J. D. (2006), Prosodic influence on syntactic judgments, in G. Fanselow, C. Fery, M. Schlesewsky & R. Vogel, eds, *Gradience in Grammar: Generative Perspectives*, Oxford, Oxford University Press.

[60] Kruschke, J. K. (1992), Alcove: An exemplar-based connectionist model of category learning, *Psychological Review* **99**, 22 – 44.

[61] Labov, W. (1972), *Sociolinguistic Patterns*, University of Pennsylvania Press, Philadelphia.

[62] Labov, W. (1990), The intersection of sex and social class in the course of linguistic change, *Language Variation and Change* **2**, 205 – 254.

[63] Labov, W. (1998), Co-existent systems in African American Vernacular English, in S. Mufwene, J. Rickford, G. Bailey & J. Baugh, eds, *African American English: Structure, history, and use*, Routledge, New York, 110 – 153.

[64] Lakoff, G. (1973), Hedges: A study in the meaning criteria and the logic of fuzzy concepts, *Journal of Philosophical Logic*.

[65] Lux, T. (1997), The voice you hear when you read silently, in *New and Selected Poems, 1975 - 1995*, Houghton Mifflin, New York.

[66] Maclagan, M. (1999), Early New Zealand English: An acoustic study, in Linguistic Society of New Zealand, Palmerston North.

[67] Maclagan, M. & Hay, J. (2004), The rise and rise of NZE DRESS, in *Proceedings of the 10th Australian International conference on Speech Science and Technology*, Sydney, 183 – 188.

[68] Maclagan, M. & Hay, J. (2007), Getting fed up with our feet: Contrast maintenance and the New Zealand English "short" front vowel shift, *Language Variation and Change* **15**, 125.

[69] Maclagan, M. A. (2000), Where are we going in our language? New Zealand English today, *NZ Journal of Speech-Language Therapy* **53**, 14 – 20.

[70] Maclagan, M. A. & Gordon, E. (1999), Data for New Zealand social dialectology: the Canterbury Corpus, *New Zealand English Journal* **13**, 50 – 58.

[71] Maclagan, M. A., Gordon, E. & Lewis, G. (1999), Women and sounds change: conservative and innovative behaviour by the same speakers, *Language Variation and Change* **11**, 19 – 41.

[72] Manning, C. D. (2003), Probabilistic syntax, in R. Bod, J. Hay & S. Jannedy, eds, *Probabilistic linguistics*, MIT Press, Cambridge, MA, 289 –341.

[73] Manning, C. D. & Schutze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Boston, MA.

[74] McKenzie, J. (2005), But he's not supposed to see me in my weeding dress!, *New Zealand English Journal* **19**, 13 – 25.

[75] Mcrobbie-Utasi, Z. & Starks, D. (2003), Phonetic realizations of the New Zealand KIT vowel in relation to two social variables, in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona.

[76] Mendoza-Denton, N., Hay, J. & Jannedy, S. (2003), Probabilistic sociolinguistics: Beyond variable rules, in R. Bod, J. Hay & S. Jannedy, eds, *Probabilistic Linguistics*, MIT Press.

[77] Monsell, S., Doyle, M. C. & Haggard, P. N. (1989), Effects of frequency on visual word recognition tasks: Where are they?, *Journal of Experimental Psychology: General* **118**, 43 – 71.

[78] Munson, B., McDonald, E., DeBoe, N. & White, A. (2006), Acoustic and perceptual bases of judgments of women and mens sexual orientation from read speech, *Journal of Phonetics* **34**, 202 – 7240.

[79] Nagata, H. (1988), The relativity of linguistic intuition: The effect of repetition on grammaticality judgments, *Journal of Pscycholinguistic Research*, **17**(1), 1 – 17.

[80] Newmeyer, F. (2003), Grammar is grammar and usage is usage, *Language* **79**, 682 – 707.

[81] Niedzielski, N. (1999), The effect of social information on the perception of sociolinguistic variables, *Journal of Language and Social Psychology*, **18**(1), 62 – 85.

[82] Nosofsky, R. M. (1986), Attention, similarity, and the identification categorization relationship, *Journal of Experimental Psychology: General*, **115**, 39 – 57.

[83] Oldfield, R. & Wingfield, A. (1965), Response latencies in naming objects, *uarterly Journal of Experimental Psychology*, **17**, 273 – 281.

[84] Pateman, T. (1987), *Language in mind and language in society: Studies in linguistic reproduction*, Clarendon Press, Oxford.

[85] Phillips, B. S. (1998), Word frequency and lexical diffusion in english stress shifts, in R. M. Hogg & L. van Bergen, eds, *Historical Linguistics 1995: Germanic Linguistics*, Vol. 2, Benjamins, Amsterdam, 96 – 105.

[86] Phillips, B. S. (2001), Lexical diffusion, lexical frequency, and lexical analysis, in J. Bybee & P. J. Hopper, eds, *Frequency and the Emergence of Linguistic Structure*, Benjamins, Amsterdam.

[87] Pierrehumbert, J. (2001), Frequency and the emergence of linguistic structure, in J. Bybee & P. Hopper, eds, *Exemplar Dynamics: Word frequency, lenition and contrast*, John Benjamins, Amsterdam, 137 – 157.

[88] Pierrehumbert, J. (2006), The next toolkit, *Journal of Phonetics* **34**(6), 516 – 530.

[89] Purnell, T., Idsardi, W. J. & Baugh, J. (1999), Perceptual and phonetic experiments on American English dialectidentification, *Journal of Language and Social Psychology* **18**(1), 10 – 30.

[90] Quinn, H. (1995), *Variation in NZE syntax and morphology: A study of the acceptance and use of grammatical variants among Canterbury and West Coast teenagers*, MA Thesis, University of Canterbury, Christchurch.

[91] Quinn, H. (2000), Variation in New Zealand English syntax and morphology, in A. Bell & K. Kuiper, eds, *New Zealand English*, Benjamins, Amsterdam, 173 – 197.

[92] Quinn, H. (2004), Possessive have and (have) got in New Zealand English, in *NWAV 33*, University of Michigan, Ann Arbor.

[93] Quinn, H. (2005), *The distribution of pronoun case forms in English*, John Benjamins, Amsterdam & Philadelphia.

[94] Remez, R. E., Fellowes, J. M. & Rubin, P. E. (1997), Talker identification based on phonetic information, *Journal of Experimental Psychology: Human Perception and Performance* **23**, 651 – 666.

[95] Renouf, A. (1987), Corpus development, in J. Sinclair, ed., *Looking Up*, Collins ELT, London.

[96] Schütze, C. T. (1996), *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*, University of Chicago Press, Chicago.

[97] Sheffert, S. M., Pisoni, D. B., Fellows, J. M. & Remez, R. E. (2002), Learning to recognize talkers from natural, sinewave and reversed speech samples, *Journal of Experimental Psychology: Human Perception and Performance* **28**, 1447 – 1469.

[98] Strand, E. & Johnson, K. (1996), Gradient and visual speaker normalization in the perception of fricatives, in D. Gibbon, ed., *Natural language processing and speech technology: results of the 3rd KONVENS conference*, Mouton de Gruyter, Berlin, 14 – 26.

[99] Strand, E. A. (1999), Uncovering the role of gender stereotypes in speech perception, *Journal of Language and Social Psychology* **18**(1), 86 – 99.

[100] Tagliamonte, S. A. (2001), Come/came variation in English dialects, *American Speech* **76**(1), 42 – 61.

[101] Team (2004), *R: A language and environment for statistical computing*.

[102] Trudgill, P. (1974), *The Social Differentiation of English in Norwich*, Cambridge University Press, Cambridge.

[103] Trudgill, P. (1988), Norwich revisited: Recent linguistic changes in an English urban dialect, *English World-Wide* **9**, 33 – 49.

[104] Trudgill, P., Gordon, E. & Lewis, G. (1997), New dialect formation and southern hemisphere English: The New Zealand short front vowels, *Journal of Sociolinguistics* **2**(1), 35 – 52.

[105] Van Lancker, D., Kreiman, J. & Wickens, T. D. (1985), Familiar voice recognition: Patterns and parameters. Part I. recognition of backwards voices, *Journal of Phonetics* **13**, 19 – 38.

[106] Vetter, H.J., Volovecky, J., & Howell, R.W. (1979), Judgments of grammaticalness: A partial replication and extension, *Journal of Psycholinguistic Research*, **8**, 567 – 583.

[107] Vitevitch, M. S. & Luce, P. (1998), When words compete: Levels of processing in spoken word perception, *Psychological Science* **9**, 325 – 329.

[108] Vitevitch, M. S. & Luce, P. (1999), Probabilistic phonotactics and spoken word recognition, *Journal of Memory & Language* **40**, 374 – 408.

[109] Vogel, R. (2006), Degraded acceptability and markedness in syntax, and the stochastic interpretation of optimality theory, in G. Fanselow, C. Fery, M. Schlesewsky & R. Vogel, eds, *Gradience in Grammar: Generative Perspectives*, Oxford, Oxford University Press.

[110] Walker, A. (2005*a*), New Zealanders would of not said that - changes in modal and perfective have constructions in New Zealand English. Paper presented at the 2nd International Postgraduate Linguistics Conference, Victoria University, Wellington.

[111] Walker, A. (2005*b*), *Well, when you say it: The effect of speaker identity on the grammaticality ratings of morpho-syntactic constructions in New Zealand English*, Honours Extended Essay.

[112] Walker, A. (2007), The effect of phonetic detail on perceived speaker age and social class, in J. Trouvain & W. J. Barry, eds, *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrucken.

[113] Watson, C., Maclagan, M. & Harrington, J. (1998), Acoustic evidence for vowel change in New Zealand English, in *The proceedings of Laboratory Phonology VI*, York.

[114] Woods, E. (2003), TH-fronting: the substitution of f/v for T/D in New Zealand English, *New Zealand English Journal* **17**, 50 – 56.

[115] Woods, N. (2000), New Zealand English across the generations: an analysis of selected vowel and consonant variables, in A. Bell & K. Kuiper, eds, *New Zealand English*, Victoria University Press and John Benjamins, Wellington and Amsterdam, 84 – 110.

[116] Woods, N. J. (1997), The formation and development of New Zealand English: Interaction of gender-related variation and linguistic change, *Journal of Sociolinguistics* **1**(1), 495– 125.

[117] Zuraw, K. (2003), Probability in language change, in R. Bod, J. Hay & S. Jannedy, eds, *Probabilistic Linguistics*, MIT Press, 139 – 176.